Aus dem Institut für Medizinische Genetik und Humangenetik
der Medizinischen Fakultät Charité - Universitätsmedizin Berlin

DISSERTATION

Evaluation der diagnostischen Genauigkeit eines Systems zur computergestützten fazialen Phänotypisierung syndromaler Patientinnen und Patienten

zur Erlangung des akademischen Grades
Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

von

Nurulhuda Hajjir aus Berlin

Datum der Promotion: 25.06.2023

Inhaltsverzeichnis

1 Abstracts

1.1 Abstract (Englisch)

Background

In syndromology, computer-aided facial analysis of patients with facial dysmorphisms has become a significant tool in the diagnosis of genetic syndromic disorders. Through machine learning, software such as Face2Gene, Clinical Face Phenotype Space, and FaceBase is trained in the detection of facial dysmorphic features through automated image analysis. Based on comparison to previous images, a list of differential diagnoses is presented. Given the high cost of genetic analysis and the rarity of individual genetic syndromes, automated image analysis can help clinicians shorten a diagnostic odyssey.

However, images of facially inconspicuous individuals cannot be identified as such. This study tests 1) Face2Gene's sensitivity, 2) whether Gestalt scores of syndromic faces differ significantly from those of inconspicuous faces, and 3) how the suggested differential diagnoses are distributed within the healthy control cohort (specificity of the system), and 4) whether ethnic background or gender affect Face2Gene's diagnostic accuracy.

Methods

For the syndromic cohort, we used photographs taken from patients with a total of 17 diagnosed syndromes, so that each syndrome was represented by an equal number of images. Each photograph was matched to the image of a healthy control subject, according to age, gender, and ethnicity to form the control cohort. DeepGestalt (v. 19.1.7) from FDNA was used to phenotype the cohorts.

Results

A total of 19 images per syndrome could be analyzed, i.e. both the syndromic cohort and the control cohort were composed of 323 images each. The high sensitivity of DeepGestalt was confirmed. In 91% of the cases, the correct diagnosis appears among the 10 highest ranked of the differential diagnoses suggested by Face2gene. Scores within the syndromic cohort were higher than scores within the facially unremarkable control cohort. A specially trained classifier achieves better results.

Conclusions

This is the first study to compare Face2Gene's results on images of inconspicuous and syndromic individuals, analyzing the specificity of DeepGestalt. Until now, detection of healthy individuals by image analysis with FDNA's DeepGestalt is not possible.

Nevertheless, significant differences were found in the scoring of healthy individuals and individuals with dysmorphic facial features. It does not appear impossible that future programs could be trained to label an individual as "inconspicuous" or healthy. For pediatricians in a clinic without trained geneticists, not only the prioritization of potentially present diseases is helpful, but also the labeling of a patient as facially inconspicuous in order to direct the focus of diagnostics to other causes of disease.

1.2 Abstract (Deutsch)

Hintergrund

In der Syndromologie ist die computergestützte Gesichtsanalyse von Patient*innen mit fazialen Dysmorphien zu einem bedeutenden Instrument in der Diagnostik genetisch-syndromaler Erkrankungen geworden. Durch maschinelles Lernen werden Softwareprogramme wie *Face2Gene*, *Clinical Face Phenotype Space* und *FaceBase* in der Erkennung dysmorpher Gesichtszüge durch automatisierte Bildanalyse trainiert. Abhängig von der Übereinstimmung zwischen dem Bild einer Person und dem System zugrunde liegenden Bildern anderer Betroffener wird eine Liste von Differentialdiagnosen präsentiert. Angesichts der hohen Kosten genetischer Analysen und der Seltenheit einzelner genetischer Syndrome kann die automatisierte Bildanalyse Kliniker*innen helfen, eine diagnostische Odyssee zu verkürzen.

Face2Gene ist allerdings so angelegt, dass jedem Bild eine Liste von Differentialdiagnosen zugeordnet wird. Bilder von fazial unauffälligen Personen können also nicht als solche erkannt werden. Diese Studie prüft 1) Face2Gene's Sensitivität, 2) ob sich die Gestalt Scores für syndromale Gesichter von denen für unauffällige Gesichter signifikant unterscheiden und 3) wie sich die vorgeschlagenen Differentialdiagnosen innerhalb der gesunden Kontrollkohorte verteilen (Spezifität des Systems) und 4) ob der ethnische Hintergrund bzw. das Geschlecht Face2Gene's diagnostische Genauigkeit beeinflussen.

Methoden

Für die Syndrom-Kohorte verwendeten wir Fotos, die von Patient*innen mit insgesamt 17 diagnostizierten Syndromen stammten, sodass jedes Syndrom durch eine gleiche Anzahl an Bildern repräsentiert wurde. Jedes Foto wurde dem Bild einer gesunden Kontrollperson zugeordnet, wobei sich Alter, Geschlecht und Ethnizität entsprachen, um die Kontrollkohorte zu bilden. Zur Phänotypisierung der Kohorten wurde DeepGestalt (v. 19.1.7) von FDNA verwendet.

Ergebnisse

Insgesamt konnten pro Syndrom 19 Bilder analysiert werden, d.h. sowohl die Syndrom-Kohorte als auch die Kontrollkohorte setzte sich jeweils aus 323 Bildern zusammen. In 91% der Fälle zeigt sich die richtige Diagnose unter den 10 höchst Platzierten der von Face2gene vorgeschlagenen Differentialdiagnosen. Auch die Scores innerhalb der Syndrom-Kohorte zeigen höhere Werte als die Scores innerhalb der fazial unauffälligen Kontrollkohorte. Ein speziell dafür trainierter Klassifikator erzielt bessere Ergebnisse.

Schlussfolgerungen

Dies ist die erste Studie, die Face2Genes Ergebnisse an Bildern unauffälliger und syndromaler Personen vergleicht und damit die Spezifität von DeepGestalt analysiert. Bis jetzt ist eine Erkennung gesunder Individuen durch Bildanalyse mit FDNA's DeepGestalt nicht möglich. Dennoch fanden sich signifikante Unterschiede im Scoring von Gesunden und Personen mit dysmorphen Gesichtszügen. Es erscheint nicht ausgeschlossen, dass künftige Programme darin trainiert werden können, eine Person als "unauffällig" oder gesund zu kennzeichnen. Für Kinderärzt*innen in einer Klinik ohne spezialisierte Genetik ist nicht nur die Priorisierung potentiell vorliegender Erkrankungen hilfreich, sondern auch die Kennzeichnung einer Person als fazial unauffällig, um den Fokus der Diagnostik auf andere Genesen zu richten.

2 Einführung

Als Teilbereich der Medizinischen Genetik liegt der Schwerpunkt der klinischen Syndromologie auf dem Diagnostizieren genetisch bedingter Syndrome. Die Dysmorphologie ist die Kunst und Wissenschaft diese genetisch bedingten Syndrome anhand von Kombinationen spezifischer Merkmale und Fehlbildungen zu erkennen. Für die Syndromologie von besonderer Bedeutung ist die klinische Interpretation spezieller Gesichtsmerkmale und deren Zuordnung zu einer genetisch-syndromalen (Verdachts-)Diagnose: die faziale Phänotypisierung.

Für Erkrankungen mit vergleichsweise hoher Inzidenz und ausgeprägten charakteristischen fazialen Stigmata gelingt eine faziale Phänotypisierung relativ einfach und routiniert. Vom Zeitpunkt des Auftretens erster Symptome bzw. der Feststellung vorhandener Dysmorphien bis zur endgültigen Diagnosefindung vergeht hier unerheblich viel Zeit. Die Verdachtsdiagnose einer Trisomie 21 etwa wird häufig bereits bei der ersten Vorsorgeuntersuchung (U1) im Kreißsaal gestellt. Je seltener eine Erkrankung jedoch ist und je unspezifischer der Phänotyp, desto schwieriger ist allerdings für die behandelnden Ärzt*innen die Diagnostik. Oft vergehen Jahre bis eine zugrundeliegende Mutation oder Diagnose benannt wird.
Weiterhin beeinflussen die Verfügbarkeit von erfahrenen klinischen Genetiker*innen, die Prävalenz bestimmter Erkrankungen, sowie die Kosten(-übernahme) von molekulargenetischen und weiteren Untersuchungen die Odyssee an diagnostischen Schritten.

In jüngster Zeit wurden mithilfe der Methoden des maschinellen Lernens und der digitalen Bilderkennung Programme zur computer-gestützten fazialen Phänotypisierung entwickelt. Diese sollen als Entscheidungsunterstützungssysteme den diagnostischen Prozess in der klinischen Genetik verbessern, aber auch zu Forschungszwecken, z.B. zur Definition neuer Syndrome, verwendet werden können. Über den klinischen Nutzen und die Genauigkeit dieser Verfahren ist bisher allerdings wenig bekannt.
Beispiele für solche Software sind die Technologien von Face2Gene (Gurovich et al. 2019), vom Clinical Face Phenotype Space (Ferry et al. 2014) und von FaceBase

(Hallgrímsson et al. 2020). Face2Gene ist das am meisten genutzte unter diesen Systemen. Mithilfe eines tiefen neuronalen Netzwerks (DeepGestalt), das an mehr als 17.000 Fotografien von Patient*innen mit einem genetischen Syndrom trainiert wurde, kann das Programm verwendet werden, um anhand gewöhnlicher Frontalaufnahmen von Patientengesichtern eine Liste möglicher Verdachtsdiagnosen zu erstellen. DeepGestalt vergleicht dazu automatisch die Gesichtszüge des_r abgebildeten Patienten_in mit Bildern von Menschen mit dem System bekannten potentiellen Differentialdiagnosen und gibt als Ergebnis eine Liste dieser Verdachtsdiagnosen, die anhand der Ähnlichkeit sortiert ist, aus.

Im Rahmen dieser Promotionsarbeit habe ich in verschiedenen Studien zur Evaluation von DeepGestalt gearbeitet.

In *Advances in computer-assisted syndrome recognition by the example of inborn errors of metabolism* prüften meine Kolleg*innen und ich das Ausmaß möglicher Störfaktoren auf die Sensitivität von DeepGestalt bei hereditären Stoffwechselerkrankungen. Dabei wurden die Analyseergebnisse verschiedener Kohorten hinsichtlich der Merkmale Kohortengröße, Geschlecht, Ethnizität und Alter verglichen. Die Faktoren Geschlecht und Ethnizität beeinflussten das Ergebnis nicht. Insgesamt zeigte sich eine hohe Sensitivität und Unterscheidungskraft bei phänotypisch ähnlichen Syndromen, wie z.B. Mukopolysaccharidose Typ I und Typ II (Jean T. Pantel et al. 2018).

Eine große klinische Ähnlichkeit zeigen auch die verschiedenen Defekte der Biosynthese des Glycosylphosphatidylinositols (GPI), welche die unterschiedlichen Formen der Hyperphosphatasie mit geistiger Entwicklungsverzögerung (Hyperphosphatasia with Mental Retardation Syndrome, HPMRS), die auch als Mabry-Syndrom bekannt sind, verursachen. In *Characterization of glycosylphosphatidylinositol biosynthesis defects by clinical features, flow cytometry, and automated image analysis* prüften wir die unterschiedlichen Mutationen, die dem Mabry-Syndrom ursächlichen GPI-Synthesedefekten zu Grunde liegen, hinsichtlich ihrer biochemischen Pathophysiologie und ihrer klinischen Ausprägung (Knaus et al. 2018). Dabei zeigt sich, dass eine Klassifizierung des zugrundeliegenden Gens mit

beachtlicher Sensitivität durch eine computer-gestützte faziale Phänotypisierung mittels DeepGestalt gelingt.

In *PEDIA: Prioritization of Exome Data by Image Analysis* zeigten meine Kolleg*innen und ich exemplarisch, wie das neuronale Netzwerk „DeepGestalt“ von Face2Gene zur Verbesserung der Analyse von panexomischen Daten in der klinischen Genetik verwendet werden kann. Die Studie schloss 679 Proband*innen mit 105 unterschiedlichen molekulargenetisch gesicherten Syndromen ein. Mithilfe einer Support Vektor Maschine wurde dabei ein von DeepGestalt errechneter Wert zur Quantifizierung der Ähnlichkeit eines Gesichts mit den typischen fazialen Merkmalen eines Syndroms, der Gestalt Score, mit anderen bereits etablierten Maßen für die Bewertung von Sequenzvarianten (CADD Score) und Phänotyp (Phenomizer Scores) kombiniert.
In 99% der Fälle ist die richtige Diagnose unter den ersten 10 vorgeschlagenen Differentialdiagnosen, in 86-89% der Fälle ist die Diagnose auf Platz 1 (Hsieh et al. 2019). Damit zeigt *PEDIA* eine bemerkenswerte Sensitivität in der Phänotypisierung durch die Kombination der Scores aus genetischer Information (Sequenzvariante), phänotypischer Information (standardisierte Beschreibung durch Human Phenotype Ontology-Terminologie) und Bildinformation.
Für die Platzierung auf Rang 1 gelingt eine Verbesserung der Treffsicherheit von 36 bis 74 Prozent ohne Inklusion des Bildscores auf 86 bis 89 Prozent durch die Inklusion des Bildscores.

Auch zeigten wir durch den Vergleich von Patient*innen mit Fanconi Anämie (FA) und Microcephaly-short-stature-limb-abnormality-Syndrom (MISSLA-Syndrom), wie eine Unterscheidbarkeit zwischen beiden klinisch ähnlichen Syndromen durch Gesichtserkennung möglich ist, auch wenn die Fanconi Anämie mit deutlich weniger fazialer Dysmorphie einhergeht. Eine Fehldiagnose von Patient*innen mit MISSLA - Syndrom ist aufgrund des überlappenden Phänotyps zur Fanconi Anämie beschrieben. Die Erkennung des MISSLA-Syndroms erfolgt mit einer Sensitivität von 84%, und nur 6% der MISSLA-Fälle sind fehlerhaft als Fanconi Anämie diagnostiziert worden (Danyel et al. 2019) Auch die Treffsicherheit für die Diagnose der Fanconi Anämie ist vergleichsweise hoch und zeigt, dass betroffene Patient*innen offenbar eine spezifische Facies haben.

Daraus ergeben sich relevante klinische und wissenschaftliche Fragen. Durch die App Face2Gene CLINIC soll beispielsweise klinisch tätigen Ärzt*innen eine Möglichkeit geboten werden, ihren differentialdiagnostischen Ansatz in der Diagnostik genetisch-syndromaler Erkrankungen durch die Verwendung einer automatisierten fazialen Phänotypisierung zu erweitern und die genetische Diagnostik zu priorisieren, indem eine gezielte Durchführung von Gen-Panels für spezifische Syndrome oder spezifischer Gentests anstelle einer Exomanalyse erwogen wird.

Krankheiten mit hoher morphologischer Variabilität und mit niedriger Inzidenz stellen eine Herausforderung im klinischen Alltag dar. Gerade für sie erscheint der Einsatz von Technologien wie Face2Gene daher attraktiv. Sie versprechen eine ebenso präzise wie objektive Analyse der klinisch relevanten Gesichtszüge. Allerdings sind - auch wenn sie automatisiert und nach eindeutig definierten mathematischen Modellen ablaufen - die Verfahren des maschinellen Lernens nicht absolut objektiv. Ihre Ergebnisse hängen wesentlich von der Zusammenstellung des Datensatzes ab, der zum Training eines Algorithmus verwendet wurde. Voraussetzung für das *Erkennen* eines Syndroms in einem dysmorphen Gesicht ist für die Software - ebenso wie für einen menschlichen Dysmorphologen - das *Kennen* der entsprechenden Krankheit. D.h., dass Face2Gene nur solche Syndrome vorschlagen kann, zu denen auch eine ausreichende Zahl an Beispielbildern in seinem Trainingsdatensatz vorkommt und dass die Fähigkeit des Systems, diese Syndrome zu detektieren, mit der Zahl der entsprechenden Bilder im Trainingsdatensatz zunimmt. Damit einher gehen mögliche Schwächen in der computer-gestützten Erkennung syndromaler Erkrankungen. So ist z.B. unklar, wie gut ein solches System bei Angehörigen verschiedener ethnischer Gruppen (Menschen afrikanischer, asiatischer, europäischer Herkunft, etc.) funktioniert, da eine mögliche Beeinflussung durch spezifische ethnische Unterschiede der Gesichtsmorphologie im Phänotyp denkbar ist und der Trainingsdatensatz nicht zwingend für alle diese Gruppen repräsentativ ist. Gleiches gilt für Bilder von Menschen verschiedenen Alters und für mögliche Sensitivitätsunterschiede aufgrund des Geschlechts. Unterschiede in der Repräsentanz bestimmter Syndrome, Ethnien bzw. Altersgruppen im Trainingsdatensatz von DeepGestalt sind anzunehmen, da die Prävalenz der einzelnen Krankheiten in der Bevölkerung in Abhängigkeit von diesen

Variablen schwankt. Auch die Bildqualität (Tragen einer Brille, Auflösung, Lichteffekte etc.) könnte die Genauigkeit der Analyse beeinflussen.
Dies sind relevante Forschungsfragen, aus denen sich Ansätze ergeben, mögliche Lücken in der Einsatzfähigkeit durch die Erweiterung und Verbesserung des Trainingssets und die Entwicklung zusätzlicher Scores zu schließen.

Es ist nicht ohne Weiteres auszuschließen, dass auch die Fähigkeit, gesunde Proband*innen als solche zu erkennen, sie also *keiner* syndromalen Krankheit zuzuordnen, in einem bioinformatischen Modell erfasst werden könnte. DeepGestalt wurde allerdings nicht für diese Aufgabe trainiert und kann daher nicht direkt zu diesem Zweck verwendet werden. In den oben genannten Arbeiten zeigte sich, dass die Zuordnung von Differentialdiagnosen in Kohorten unauffälliger Kontrollbilder gewisse Wiederholungsmuster aufweisen. Hier sind die erfahrenen klinischen Genetiker*innen möglicherweise weniger fehleranfällig.
Ein Modell zur Erkennung gesunder Proband*innen ohne kraniofaziale Dysmorphien sollte perspektivisch auch die Qualität der Differentialdiagnostik in Kohorten mit kraniofazialen Dysmorphien verbessern. Zur Prüfung der Spezifität für Gesunde wurde eine entsprechende Kontrollkohorte zusammengestellt und Face2Gene's Ergebnisse systematisch ausgewertet.

In meiner Publikation *„Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals with and without a Genetic Syndrome: Diagnostic Accuracy Study"* haben meine Kolleg*innen und ich so zuletzt die Genauigkeit der automatisierten Gesichtserkennung bei Patient*innen mit fazialen Dysmorphien und einer gesunden Kontrollkohorte bestimmt und gezeigt, dass der Software *DeepGestalt* eine Differenzierung grob gelingt, eine trennscharfe Klassifizierung eines Gesichts als unauffällige Facies allerdings noch nicht möglich ist.

3 Material und Methodik

3.1 Bilder von symdromalen Proband*innen

Für die Syndrom-Kohorte wurden Portraitfotos von Patient*innen mit insgesamt 17 klinisch beziehungsweise molekulargenetisch gesicherten Diagnosen verwendet. Die Zahl der pro Diagnose gesammelten Bilder sollte jeweils gleich groß sein. Die

Patientenbilder wurden aus Publikationen entnommen oder zum Zwecke dieser Studie erstellt, nachdem das informierte, schriftliche Einverständnis eingeholt wurde.
Die Auswahl der Syndrome dieser Studie basiert auf den insgesamt 201 verschiedenen Diagnosen, die Face2Gene Clinic in der gesunden Kontrollkohorte, die wir in Danyel et al. beschrieben haben, vorgeschlagen hat. Es wurden gezielt Syndrome mit unterschiedlich häufiger Nennung unter den möglichen Differentialdiagnosen ausgewählt. Die relativen Häufigkeiten reichten dabei von 1% bis 76%. Die untersuchten Syndrome waren: Fragiles-X-Syndrom (MIM:#300624), Angelman-Syndrom (MIM:#105830), Rett-Syndrom (MIM:#312750), Phelan-Mcdermid-Syndrom (MIM:#606232), Klinefelter-Syndrom, Beckwith-Wiedemann-Syndrom (MIM:#130650), 22q11.2-Deletions-Syndrom (MIM:#611867), Sotos-Syndrom (MIM:#117550), Noonan-Syndrom (MIM:PS163950), Loeys-Dietz-Syndrom (MIM:PS609192), Williams-Beuren-Syndrom (MIM:#194050), Rubinstein-Taybi-Syndrom (MIM:PS180849), Achondroplasie (MIM:#100800), Wolf-Hirschhorn-Syndrom (MIM:#194190), Pallister-Killian-Syndrom (MIM:#601803), und Treacher-Collins-Syndrom (MIM:PS154500). Als ein Beispiel für eine Diagnose, die nicht unter den falsch-positiven Nennungen der gesunden Kontrollkohorte von Danyel et al. zu finden war, wurde das Apert-Syndrom verwendet (siehe Figure 1, Pantel et al. 2020)

3.2 Bilder gesunder Kontrollpersonen

Um eine Kohorte negativer (d.h. unauffälliger) Kontrollbilder zu bilden, wurde für jedes Foto aus der Syndrom-Kohorte ein nach Alter, ethnischem Hintergrund und Geschlecht vergleichbares Bild verwendet. Die Abwesenheit klinisch-relevanter kraniofazialer Auffälligkeiten wurde durch Ärzt*innen mit Erfahrung im Bereich der Humangenetik beurteilt.

3.3 Faziale Phänotypisierung

Die digitale Phänotypisierung aller Fotos erfolgte mit DeepGestalt v. 19.1.7 ausschließlich über die eingepflegten Bilder. Es wurden weder klinische Informationen in Form von HPO-Terms noch molekulargenetische Informationen oder Diagnosen an DeepGestalt übergeben.

3.4 Statistische Analyse

Für die Erstellung der Klassifikation “auffällig” versus “ unauffällig” auf Basis der von DeepGestalt ausgegebenen Gestalt-Scores wurden lineare Support-Vector-Maschinen (SVM) in Scikit-Learn (Version 0.21.3) mit Python 3.7 in einem leave-one-out Schema trainiert.

Die diagnostische Genauigkeit wurde mittels der Fläche unter der Isosensitivitätskurve (ROC-Kurve) bestimmt. Ein idealer Test erzielt den Maximalwert von einer *area under the ROC-curve* (AUROC) von 1. Ein unbrauchbarer Test, der lediglich nach dem Zufallsprinzip klassifiziert, erzielt einen Wert von 0,5. Eine komplette Fehlklassifikation (alle Positiven werden negativ, alle Negativen werden als positiv klassifiziert) erzielt eine AUROC von 0. Unterschiede zwischen den ROC-Kurven wurden mittels DeLong-Test geprüft. Unterschiede in der Verteilung der Scores wurden mittels Welch-Test geprüft. Die Daten wurden auch nach den Variablen Ethnie und Geschlecht entsprechend statistisch ausgewertet. Für die Analyse nach dem Geschlecht wurden allerdings das Rett-Syndrom und das Fragile-X-Syndrom ausgeschlossen, denn diese Syndrome folgen X-chromosomalen Erbgängen. Die Frequenz der Betroffenen ist daher stark abhängig vom Geschlecht.

4 Ergebnisse

4.1 Bilder und Matching

Für die Studie wurden insgesamt 646 Bilder verwendet, wovon 323 Bilder Menschen ohne faziale Dysmorphien darstellen und die Kontrollkohorte bilden, die andere Hälfte zeigt Menschen mit fazialen Dysmorphien und bildet die Syndrom-Kohorte. Hierbei sind je 19 Bilder für die 17 verschiedenen genetische Syndrome vertreten. Der Anteil der weißen Menschen beträgt 83%. Das Verhältnis weiblich zu männlich beträgt etwa 1:1 (160/161). Bei zwei Bildern war eine Geschlechtszuordnung nicht möglich.

4.2 Sensitivität von DeepGestalt

Die Zuordnung der Fotos syndromaler Proband*innen zur richtigen Diagnose durch DeepGestalt erfolgte - wie erwartet - mit hoher Sensitivität. Die richtige Diagnose erscheint in durchschnittlich 91% der Fälle in den Top 10 der Vorschläge, in 61% der Fälle sogar auf Rang 1 (Figure 2b, (Pantel et al. 2020). Die hohe Detektionsrate lässt

sich sowohl bei Bildern von Patient*innen europäischer Herkunft (Top-10-Sensitivität 90%) als auch bei Bildern von Patient*innen mit einer anderen ethnischen Herkunft (Top-10-Sensitivität 97%) feststellen. Weiterhin prüften wir die Sensitivität geschlechtsspezifisch. Es zeigt sich kein relevanter Unterschied in der Sensitivität zwischen den Ergebnissen der weiblichen und der männlichen Kohorte.
Fragiles-X- Syndrom, Noonan-Syndrom, Phelan-McDermid-Syndrom, Rett-Syndrom, Sotos-Syndrom, Treacher-Collins und Williams-Beuren-Syndrom erscheinen in der Phänotypisierung in jeweils allen 19 Fällen unter den Top 10. DeepGestalt weist für diese Syndrome folglich eine Sensitivität von 100% auf. Selbst für das Loeys-Dietz-Syndrom, für welches das System die niedrigste Detektionsrate unter den getesteten Syndromen zeigte, erreicht es immerhin eine Top-10-Sensitivität von 74%.

4.3 Spezifität von DeepGestalt

Die Verteilung der Diagnosevorschläge innerhalb der gesunden Kohorte zeigt ein bemerkenswertes Muster. Unter den 238 in der Kontrollkohorte vorgeschlagenen Syndromen sind beispielsweise Fragiles-X-Syndrom, Angelman-Syndrom, Rett-Syndrom und Klinefelter-Syndrom weit überzufällig häufig vertreten. So findet sich in mehr als 80% der Fälle Fragiles-X-Syndrom in der Liste der jeweils 30 vorgeschlagenen Differentialdiagnosen (Figure 2a, Pantel et al 2020).

Neben der Rangliste der Top 30 ist es wichtig zu berücksichtigen, wie hoch die einzelnen Scores sind: es ist anzunehmen, dass die Zuordnung einer Diagnose zu einem hohen Gestalt-Score einen höheren positiv-prädiktiven Wert hat als die weiteren Diagnosen in der Rangliste mit niedrigeren Scores. Auch die Priorisierung einer Diagnose auf Rang 1 könnte - bei niedrigem Gestalt-Score - trotzdem unwahrscheinlich sein. Der durchschnittliche Erstrang-Score in der Syndrom-Kohorte betrug 0,47, der maximale Erstrang-Score lag bei 1,0. In der Kontrollkohorte lag der durchschnittliche Erstrang-Score bei 0,27, der maximale Erstrang-Score betrug 0,85. Mit einer Fläche unter der ROC-Kurve (AUROC) von 0,72 zeigten die höchsten Gestalt-Scores von Syndrom-Kohorte und Kontrollkohorte eine Unterscheidbarkeit. Die Unterschiede in der Verteilung der Scores waren dabei statistisch signifikant ($P<.001$). Wir verglichen außerdem die Ergebnisse geschlechtsspezifisch. Für die weibliche Kohorte ergab sich eine AUROC von 0,71, für die männliche Kohorte

ebenfalls eine AUROC von 0,71 (s. Abbildung 1). Die Performance von DeepGestalt unterscheidet sich also nicht in den geschlechtsspezifischen Kohorten.

Mit der SVM Klassifikation durch Stützvektormethode (*Support Vector Machine*) in die Kategorien “auffällig” und “unauffällig” ist eine noch stärkere Trennbarkeit erzielt worden (AUROC 0,89, P<.001). Auch hier lassen sich Syndrome nennen, die in allen Fällen korrekt zugeordnet wurden, im Sinne der binären Klassifikation also als syndromal erkannt wurden, nämlich Apert-Syndrom, Wolf-Hirschhorn-Syndrom und Williams-Beuren-Syndrom. Das Klinefelter-Syndrom liefert mit nur 7 richtigen Zuordnungen das schlechteste Ergebnis.
Insgesamt ist die AUROC der SVM-Klassifikation höher als die AUROC der Gestalt-Scores und zeigt damit eine bessere Differenzierung. Das Ergebnis ist umso besser, je mehr potentielle Diagnosen, d.h. je mehr der jeweils höchsten Gestalt-Scores in die Klassifizierung der SVM einfließen. Weiterhin erfolgte auch hier eine geschlechtsspezifische Klassifizierung. Für die weibliche Kohorte ergab sich eine AUROC von 0,93, für die männliche Kohorte eine AUROC von 0,86 (s. Abbildung 1).

Die geschlechtsspezifischen Kurven von DeepGestalt und SVM-Klassifikation wurden im letzten Schritt verglichen und zeigten für die weibliche, für die männliche und für die gemischte Kohorte einen p-Wert von <0,001 und zeigen damit eine Trennbarkeit (s. Abbildung 1).

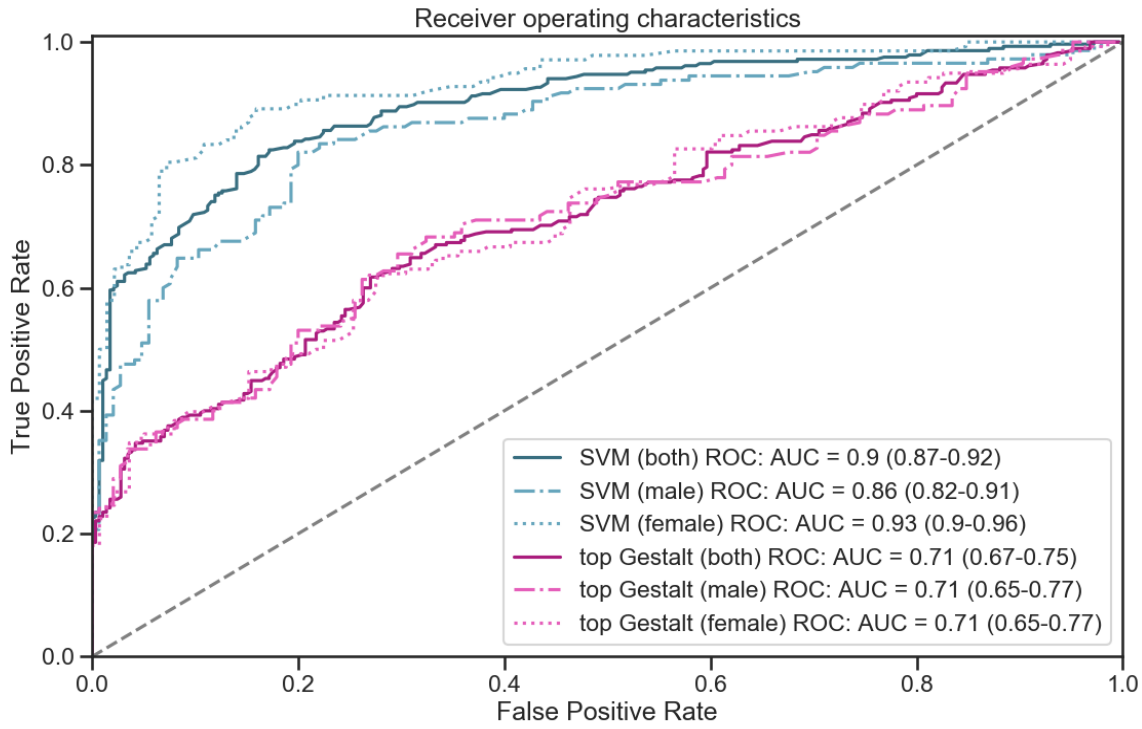


Abbildung 1: *Isosensitivitätskurven der Bildklassifikationen durch DeepGestalt (lila) und durch die SVM (türkis)*
Pantel et al. 2020, Figure 3

4.4 Bias

Die Ergebnisse der Bildanalysen von Personen europäischer Herkunft sind vergleichbar mit denen der Bildanalysen von Menschen anderer Herkunft. Die diagnostische Trennschärfe bezüglich der Differenzierung in auffällig versus unauffällig zeigt sich in der Kohorte der Menschen mit einem europäischen Hintergrund zwar besser als in der Kohorte mit anderen Ethnien, allerdings ist erstere Kohorte auch quantitativ größer - der Anteil von Bildern weißer Personen beträgt 84% (272 von 323 Bildern). Ein Rückschluss auf statistisch signifikante Leistungsunterschiede in Abhängigkeit vom ethnischen Hintergrund der Abgebildeten ist bei vorliegender Ausgangsverteilung daher nicht möglich. Die Geschlechtsverteilung ist mit einem Verhältnis von 160 zu 161 allerdings fast gleichmäßig. Hier zeigt sich keine relevante Differenz in der Performance.

In der weiblichen Syndrom-Kohorte betrug der maximale Score 0,99, der durchschnittliche Score 0,48 und der minimale Score 0,085, in der entsprechenden Kontrollkohorte betrug der maximale Score 0,84, der durchschnittliche: Score 0,27 und der minimale 0,08. Für die männlichen Kohorte ergeben sich in der Syndrom-Kohorte ein maximaler Score 0.99, ein durchschnittlicher Score von 0,47 und ein minimaler

Score von 0,08, für die gesunde Kontrollkohorte ergaben sich ein maximaler Score von 0.79, ein Durchschnittsscore von 0.27 und ein minimaler Score von 0.06. Das Geschlecht des_r Patienten_in sehen wir daher nicht als Bias.

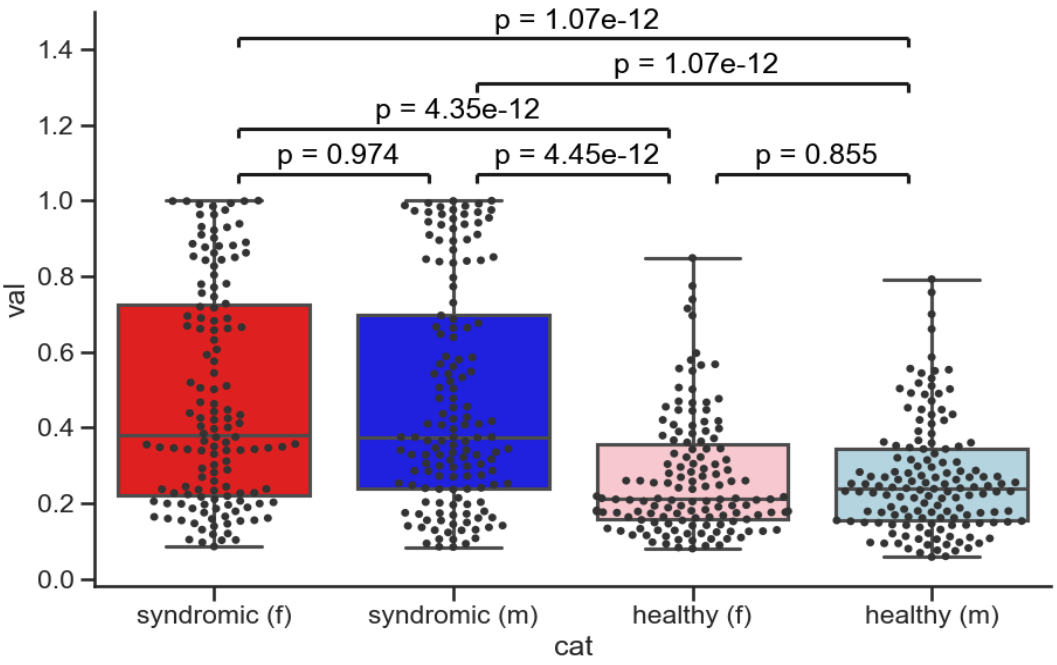


Abbildung 2: *Verteilung der maximalen Gestalt scores bei Bildern von syndromalen und unauffälligen Personen im Geschlechtervergleich. m = männlich, f = weiblich*
Dargestellt werden die Daten von Pantel et al .2020.

5 Diskussion

Die hohe Sensitivität der Testkohorte bestätigt die bereits in der *PEDIA-Studie* präsentierten Ergebnisse, die in einer älteren technischen Version von *DeepGestalt* zustande gekommen sind (Hsieh et al. 2019). In der PEDIA-Studie haben wir gezeigt, dass durch den Einschluss des Gestalt-Scores in die Priorisierung der Differentialdiagnosen die Treffsicherheit erheblich ansteigt. Die Sensitivität von *DeepGestalt* und *PEDIA* steigt außerdem mit dem wachsenden Trainingsdatensatz. Eine niedrige Treffsicherheit bei der automatisierten Identifizierung vorliegender genetischer Krankheiten könnte zum Einen auf die geringe Prävalenz des jeweiligen Syndroms und zum Anderen auf die relative Neubeschreibung eines Syndroms zurückgeführt werden. In beiden Fällen stünde der Software in der Konsequenz ein vergleichsweise kleiner Satz an Patientenbildern für das Trainieren der Syndrommodelle zur Verfügung. Dementsprechend könnte die Genauigkeit von

DeepGestalt variieren. Dies ist eine mögliche Erklärung für die Syndrom-abhängige Spezifität von *DeepGestalt*.

Auch eine genetische Heterogenität von Syndromen erschwert eine Gen-spezifische Phänotypisierung durch *DeepGestalt,* da die jeweiligen Subtypen eines solchen Syndroms mitunter im Trainingsdatensatz unter einer Klasse zusammengefasst werden (z.B. Noonan-Syndrom, Kabuki-Syndrom).

Die Größe des Trainingssets als Störfaktor ist allerdings ein veränderlicher Parameter und unterliegt einer positiven Dynamik. Die neueren Versionen von DeepGestalt, die mit größerem Datensatz arbeiten, erzielen eine höhere Sensitivität. So zeigte sich in dieser Arbeit eine Top 10 Sensitivität von 91% für die gesamte Kohorte von 323 Patient*innen. In der Studie von Hsieh et al, die mit einer früheren Version von DeepGestalt arbeitete, zeigt sich eine Top 10 Sensitivität von 48,6% für die Kohorte von 679 Patient*innen. Wie die Genauigkeit von DeepGestalt, soll auch die diagnostische Genauigkeit von *PEDIA* durch die Vergrößerung des Datensatzes stetig verbessert werden.

Neben der Größe des Datensatzes ist auch die Zusammenführung verschiedener Informationen (Genotyp- und Phänotyp-basiert) für die Performance der Klassifikation entscheidend. Aufgrund der Ergebnisse der PEDIA-Studie ist nämlich anzunehmen, dass die Kombination der verschiedenen Scores aus molekulargenetischer, klinischer und bildbasierter Information eine bessere Trennschärfe erzielt als die Priorisierung allein anhand der bildbasierten oder der Genotyp-basierten Scores.

Eine Kombination der Scores ist insbesondere dann signifikant, wenn entweder der Phänotyp oder der Genotyp für die richtige Diagnose wegweisend sind; Krankheiten mit hoher klinischer Variabilität und bekannter Pathogenität der Genvariante sind Beispiele hierfür. Dieser Logik folgend profitieren Fälle mit pathognomonischem Phänotyp und genetischer Heterogenität auch davon. Durch den Einschluss von spezifischen klinischen Merkmalen (Bsp.: hypoplastische Nägel, fehlende Phalangen etc.) die Teil des nicht-fazialen Phänotyps sind und damit durch DeepGestalt nicht berücksichtigt werden, gelingt schließlich die sensitivste diagnostische Beurteilung.

Bei dem hier diskutierten Ansatz erfolgte die Phänotypisierung lediglich bildbasiert. Überträgt man die Ergebnisse von *PEDIA* mit kombinierten Scores auf die vorliegende Studie, so ist von einer weiteren Steigerung der Trennschärfe hinsichtlich “auffällig” und “unauffällig” auszugehen.

Eine hohe Sensitivität in der Phänotypisierung durch eine Bildanalyse mittels DeepGestalt zeigte sich auch in Knaus et al im experimentellen und klinischen Vergleich von Erkrankungen des GPI-Ankersystems. So zeigte die Kombination von Gesichtserkennung und Genotyp die höchste Sensitivität, sodass Knaus et al sogar eine neue Klassifizierung der GPI-Störungen diskutieren, die bisher lediglich anhand der Erhöhung der alkalischen Phosphatase durch die Einteilung in *Hyperphosphatasia with Mental Retardation Syndrome* (HPMRS) und *Multiple Congenital Anomalies Hypotonia Seizures Syndrome* (MCAHS) erfolgte. Da eine solch subtile Differenzierung fazialer Auffälligkeiten durch DeepGestalt möglich ist, scheint eine Unterscheidung von syndromologisch auffälligen und unauffälligen Bildern, wie sie in der aktuellen Studie demonstriert wurde, plausibel.

Die diagnostische Genauigkeit von DeepGestalt bei Bildern von Menschen europäischer Herkunft war gegenüber der Genauigkeit des Systems bei Bildern von Menschen anderer Herkunft leicht erhöht. Die Unterschiede waren jedoch nicht statistisch signifikant. Dies ist wahrscheinlich in der geringen Zahl von Bildern in der nicht-europäischen Kohorte begründet.
Allerdings ist Ethnizität als potentieller Störfaktor bereits mehrfach diskutiert worden. Lumaka et al zeigen anhand der Untersuchung von Patient*innen mit Down-Syndrom, dass die Performance der computer-gestützten Syndromerkennung bei weißen Patient*innen besser gelingt als bei afrikanischen Patient*innen (Lumaka et al. 2017), da gewisse faciale Merkmale im Kontext der ethnischen Zugehörigkeit mehr oder weniger markant sind. Ein ähnliches Ergebnis konnten wir in der Arbeit *Advances in computer-assisted syndrome recognition by the example of inborn errors of metabolism* für das Down-Syndrom reproduzieren. Im Vergleich von Patient*innen mit Mukolipidose, Mukopolysaccharidose I und II, sowie Smith-Lemli-Opitz und Nicolaides-Baraitser-Syndrom sehen wir beispielsweise keine signifikante Veränderung der Performance bei der Gegenüberstellung der Kohorten europäisch versus nicht-europäisch Herkunft (Pantel et al. 2018).
Die wahrscheinlichste Erklärung für die steigende Unabhängigkeit des Systems vom ethnischen Hintergrund in der vorliegenden Arbeit und in der Arbeit von Pantel et al, 2018 im Vergleich zur Studie von Lumaka et al. ist, dass Lumaka et al. eine sehr frühe

Version von DeepGestalt verwendet haben und inzwischen ein ethnisch diverserer Trainingsdatensatz zur Verfügung steht.
Auch gehen wir bei steigender Größe des Datensatzes von einer Verbesserung der Klassifizierung “auffällig” versus “unauffällig” aus. Das neuronale Netzwerk muss also grundsätzlich in der Lage sein, mit zumindest moderater Genauigkeit eine Facies als unauffällig zu markieren.
In dieser Arbeit fand sich keine signifikante Diskrepanz in den Ergebnissen der männlichen und weiblichen Teilkohorte. Das Geschlecht der abgebildeten Person spielt für die diagnostische Genauigkeit von DeepGestalt offenbar keine Rolle. Ähnliche Schlussfolgerungen zeigten sich bei Tests des Systems an Bildern von Proband*innen mit metabolischen Erkrankungen. (Pantel et al. 2018)

Die unterschiedliche Frequenz bestimmter Syndrome in der Liste möglicher Differentialdiagnosen wurde in einer nachfolgenden Arbeit bestätigt (Marwaha et al. 2021). Die Tatsache, dass bestimmte Syndrome von DeepGestalt überzufällig häufig vorgeschlagen werden, hat unmittelbare Konsequenzen für den klinischen Einsatz des Systems. Zur sinnvollen Interpretation der Ergebnisse von Face2Gene sollte die Falsch-Positiv-Rate der jeweiligen Differentialdiagnose genannt werden. Diagnosen mit einer hohen Falsch-Positiv-Rate sind weniger wahrscheinlich zutreffend. Solche mit einer niedrigen Falsch-Positiv-Rate werden eher erwogen. Einige der Syndrommodelle von DeepGestalt sind klinisch fragwürdig. So weist das System eine hohe Falsch-Positiv-Rate für das Klinefelter-Syndrom auf, was wenig verwunderlich erscheint, da dieses Syndrom keine spezifischen Gesichtszüge zeigt (Bird and Hurren 2016).

Der Einsatz einer automatisierten Technologie zur Gesichtserkennung wirft auch ethische Fragen auf. Wie soll mit dem Risiko der (fälschlichen) Pathologisierung von Menschen mit auffälligen Gesichtszügen umgegangen werden? Wer soll in welchem Kontext zum Einsatz der Technologie berechtigt sein? Wie kann der Einsatz der Technologie und der Umgang mit den dafür nötigen Daten kontrolliert werden? Erste Arbeiten diskutieren diese Fragen bereits (Martinez-Martin 2019; McCradden, Patel, and Chad 2021). Mit dem zunehmenden Einsatz von Bilderkennungstechnologien und Verfahren des maschinellen Lernens nicht nur in der klinischen Genetik wird künftig eine noch umfassendere ethische Auseinandersetzung damit nötig werden. Die

syndromologische Interpretation der Gesichtszüge einer Person wird mitunter von dieser oder den Eltern als unangenehm empfunden. Künftige Forschung ist nötig, um zu klären, ob eine automatisierte Gesichtsanalyse für das Arzt-Patienten-Verhältnis in der Syndromologie dienlich ist oder als negativ angesehen wird ist.

6 Schlussfolgerungen

Ob eine genetisch-syndromale Krankheit vorliegt, kann letztlich nur durch eine genetische Untersuchung sicher bestätigt werden. Häufig beschriebene Erkrankungen sind für Pädiater*innen und Humangenetiker*innen in den meisten Fällen einfacher zu diagnostizieren. Insbesondere Erkrankungen mit auffälligen phänotypischen Stigmata führen seltener zu einer diagnostischen Odyssee.
Pädiater*innen in Praxen oder kleinen Kliniken, die auf dem Gebiet der Syndromologie begrenzt erfahren sind oder in kleinen Einzugsgebieten mit anteilig wenig Erkrankten arbeiten, können von der Nutzung computer-gestützter Phänotypisierung profitieren, wenn der Verdacht einer syndromalen Krankheit besteht oder der Auftrag für eine molekulargenetische Untersuchung konkretisiert werden soll. Aber auch für ärztlich Tätige mit einem großen Erfahrungsschatz kann der Einsatz von Software wie DeepGestalt im klinischen Alltag sinnvoll sein. Eine zeitnahe Diagnosefindung gibt allen voran Eltern und Patient*innen Gewissheit und Aufschluss über die Prognose und den Verlauf der Erkrankung, sowie das Wiederholungsrisiko für weitere Kinder. Auch entscheidende therapeutische Ansätze, falls vorhanden, können früher erfolgen und gegebenenfalls einen besseren Effekt erzielen.

Aufgrund der bislang mittelgradigen Spezifität ist die Nutzung der automatisierten Bilderkennung für Nutzer*innen, die außerhalb medizinischer Professionen tätig sind, allerdings nicht sinnvoll. Für den Einsatz von DeepGestalt sollten Kliniker*innen und Genetiker*innen nach wie vor in der klinisch-diagnostischen Beurteilung fazialer Dysmorphien geschult sein. Weiterhin müssen dem_r Nutzer_in neben der Bildanalyse Ressourcen zur Ergänzung der Diagnostik zur Verfügung stehen, um eine Verdachtsdiagnose abschließend zu klären, so bedarf es etwa der Verfügbarkeit von Laboranalysen, Bildgebung und funktioneller Diagnostik. Dieser Standard ist bei breiter Verfügbarkeit etwa als App für Laien nicht zu sichern, daher halte ich eine

ausschließliche Nutzung für autorisierte Personen zum Zwecke der Diagnosefindung und zu Forschungszwecken für sinnvoll.

Die künftige Leistungsfähigkeit von DeepGestalt ist vielversprechend, muss aber durch weitere Studien begleitet werden.

7 Literaturverzeichnis

Bird, Rebecca J., and Bradley J. Hurren. 2016. "Anatomical and Clinical Aspects of Klinefelter's Syndrome." *Clinical Anatomy* 29 (5): 606–19.

Danyel, Magdalena; Cheng, Zhuo; Jung, Christine; Boschann, Felix; Pantel, Jean Tori; Hajjir, Nurulhuda; Flöttmann, Ricarda; Schulz, Solveig; Demuth, Ilja; Sheridan, Eamonn; Mundlos, Stefan; Horn, Denise; Mensah, Martin A, 2019. "Differentiation of MISSLA and Fanconi Anaemia by Computer-Aided Image Analysis and Presentation of Two Novel MISSLA Siblings." *European Journal of Human Genetics: EJHG* 27 (12): 1827–35.

Ferry, Quentin; Steinberg, Julia; Webber, Caleb; FitzPatrick, David R; Ponting, Chris P; Zisserman, Andrew; Nellåker, Christoffer. 2014. "Diagnostically Relevant Facial Gestalt Information from Ordinary Photos." *eLife* 3 (June): e02020.

Gurovich, Yaron; Hanani, Yair; Bar, Omri; Nadav, Guy; Fleischer, Nicole; Gelbman, Dekel; Basel-Salmon, Lina; Krawitz, Peter M; Kamphausen, Susanne B; Zenker, Martin; Bird, Lynne M; Gripp, Karen W, 2019. "Identifying Facial Phenotypes of Genetic Disorders Using Deep Learning." *Nature Medicine* 25 (1): 60–64.

Hallgrímsson, Benedikt; Aponte, J David; Katz, David C; Bannister, Jordan J; Riccardi, Sheri L; Mahasuwan, Nick; McInnes, Brenda L; Ferrara, Tracey M; Lipman, Danika M; Neves, Amanda B; Spitzmacher, Jared A J; Larson, Jacinda R; Bellus, Gary A; Pham, Anh M; Aboujaoude, Elias; Benke, Timothy A; Chatfield, Kathryn C; Davis, Shanlee M; Elias, Ellen R; Enzenauer, Robert W; French, Brooke M; Pickler, Laura L; Shieh, Joseph T C; Slavotinek, Anne; Harrop, A Robertson; Innes, A Micheil; McCandless, Shawn E; McCourt, Emily A; Meeks, Naomi J L; Tartaglia, Nicole R; Tsai, Anne C-H; Wyse, J Patrick H; Bernstein, Jonathan A; Sanchez-Lara, Pedro A; Forkert, Nils D; Bernier, Francois P; Spritz, Richard A; Klein, Ophir D. 2020. "Automated Syndrome Diagnosis by Three-Dimensional Facial Imaging." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 22 (10): 1682–93.

Hsieh, Tzung-Chien; Mensah, Martin A; Pantel, Jean T; Aguilar, Dione; Bar, Omri; Bayat, Allan; Becerra-Solano, Luis; Bentzen, Heidi B; Biskup, Saskia; Borisov, Oleg; Braaten, Oivind; Ciaccio, Claudia; Coutelier, Marie; Cremer, Kirsten; Danyel, Magdalena; Daschkey, Svenja; Eden, Hilda David; Devriendt, Koenraad; Wilson, Sandra; Douzgou, Sofia; Đukić, Dejan; Ehmke, Nadja; Fauth, Christine; Fischer-Zirnsak, Björn; Fleischer, Nicole; Gabriel, Heinz; Graul-Neumann, Luitgard; Gripp, Karen W; Gurovich, Yaron; Gusina, Asya; Haddad, Nechama; Hajjir, Nurulhuda; Hanani, Yair; Hertzberg, Jakob; Hoertnagel, Konstanze; Howell, Janelle; Ivanovski, Ivan; Kaindl, Angela; Kamphans, Tom; Kamphausen, Susanne; Karimov, Catherine; Kathom, Hadil; Keryan, Anna; Knaus, Alexej; Köhler,

Sebastian; Kornak, Uwe; Lavrov, Alexander; Leitheiser, Maximilian; Lyon, Gholson J; Mangold, Elisabeth; Reina, Purificación Marín; Carrascal, Antonio Martinez; Mitter, Diana; Herrador, Laura Morlan; Nadav, Guy; Nöthen, Markus; Orrico, Alfredo; Ott, Claus-Eric; Park, Kristen; Peterlin, Borut; Pölsler, Laura; Raas-Rothschild, Annick; Randolph, Linda; Revencu, Nicole; Fagerberg, Christina Ringmann; Robinson, Peter Nick; Rosnev, Stanislav; Rudnik, Sabine; Rudolf, Gorazd; Schatz, Ulrich; Schossig, Anna; Schubach, Max; Shanoon, Or; Sheridan, Eamonn; Smirin-Yosef, Pola; Spielmann, Malte; Suk, Eun-Kyung; Sznajer, Yves; Thiel, Christian T; Thiel, Gundula; Verloes, Alain; Vrecar, Irena; Wahl, Dagmar; Weber, Ingrid; Winter, Korina; Wiśniewska, Marzena; Wollnik, Bernd; Yeung, Ming W; Zhao, Max; Zhu, Na; Zschocke, Johannes; Mundlos, Stefan; Horn, Denise; Krawitz, Peter M 2019. “PEDIA: Prioritization of Exome Data by Image Analysis.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 21 (12): 2807–14.

Knaus, Alexej; Pantel, Jean Tori; Pendziwiat, Manuela; Hajjir, Nurulhuda; Zhao, Max; Hsieh, Tzung-Chien; Schubach, Max; Gurovich, Yaron; Fleischer, Nicole; Jäger, Marten; Köhler, Sebastian; Muhle, Hiltrud; Korff, Christian; Møller, Rikke S; Bayat, Allan; Calvas, Patrick; Chassaing, Nicolas; Warren, Hannah; Skinner, Steven; Louie, Raymond; Evers, Christina; Bohn, Marc; Christen, Hans-Jürgen; van den Born, Myrthe; Obersztyn, Ewa; Charzewska, Agnieszka; Endziniene, Milda; Kortüm, Fanny; Brown, Natasha; Robinson, Peter N; Schelhaas, Helenius J; Weber, Yvonne; Helbig, Ingo; Mundlos, Stefan; Horn, Denise; Krawitz, Peter M 2018. “Characterization of Glycosylphosphatidylinositol Biosynthesis Defects by Clinical Features, Flow Cytometry, and Automated Image Analysis.” *Genome Medicine* 10 (1): 3.

Lumaka, A; Cosemans, N; Lulebo Mampasi, A; Mubungu, G; Mvuama, N; Lubala, T; Mbuyi-Musanzayi, S; Breckpot, J; Holvoet, M; de Ravel, T; Van Buggenhout, G; Peeters, H; Donnai, D; Mutesa, L; Verloes, A; Lukusa Tshilobo, P; Devriendt, K, 2017. “Facial Dysmorphism Is Influenced by Ethnic Background of the Patient and of the Evaluator.” *Clinical Genetics* 92 (2): 166–71.

Martinez-Martin, Nicole. 2019. “What Are Important Ethical Implications of Using Facial Recognition Technology in Health Care?” *AMA Journal of Ethics* 21 (2): E180–87.

Marwaha, Ashish; Chitayat, David; Meyn, M Stephen; Mendoza-Londono, Roberto; Chad, Lauren, 2021. “The Point-of-Care Use of a Facial Phenotyping Tool in the Genetics Clinic: Enhancing Diagnosis and Education with Machine Learning.” *American Journal of Medical Genetics. Part A* 185 (4): 1151–58.

McCradden, Melissa D., Evani Patel, and Lauren Chad. 2021. “The Point-of-Care Use of a Facial Phenotyping Tool in the Genetics Clinic: An Ethics Tête-a-Tête.” *American Journal of Medical Genetics. Part A* 185 (2): 658–60.

Pantel, Jean Tori; Hajjir, Nurulhuda; Danyel, Magdalena; Elsner, Jonas; Abad-Perez, Angela Teresa; Hansen, Peter; Mundlos, Stefan; Spielmann, Malte; Horn, Denise; Ott, Claus-Eric; Mensah, Martin Atta, 2020. “Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals With and Without a Genetic Syndrome: Diagnostic Accuracy Study.” *Journal of Medical Internet Research* 22 (10): e19263.

Pantel, Jean T; Zhao, Max; Mensah, Martin A; Hajjir, Nurulhuda; Hsieh, Tzung-Chien; Hanani, Yair; Fleischer, Nicole; Kamphans, Tom; Mundlos, Stefan; Gurovich, Yaron; Krawitz, Peter M, 2018. “Advances in Computer-Assisted Syndrome Recognition by the Example of Inborn Errors of Metabolism.” *Journal of Inherited Metabolic Disease* 41 (3): 533–39.

8 Eidesstattliche Versicherung einschließlich Anteilserklärung

8.1 Eidesstattliche Versicherung

„Ich, Nurulhuda Hajjir, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: **Evaluation der diagnostischen Genauigkeit eines Systems zur computergestützten fazialen Phänotypisierung syndromaler Patientinnen und Patienten**
selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.
Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren/innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Ich versichere ferner, dass ich die in Zusammenarbeit mit anderen Personen generierten Daten, Datenauswertungen und Schlussfolgerungen korrekt gekennzeichnet und meinen eigenen Beitrag sowie die Beiträge anderer Personen korrekt kenntlich gemacht habe (siehe Anteilserklärung). Texte oder Textteile, die gemeinsam mit anderen erstellt oder verwendet wurden, habe ich korrekt kenntlich gemacht.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit der Erstbetreuerin angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.og) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum Unterschrift

8.2 Ausführliche Anteilserklärung an der erfolgten Publikation im Rahmen des Promotionsverfahrens

Nurulhuda Hajjir hat für die Publikation *Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals With and Without a Genetic Syndrome: Diagnostic Accuracy Study* die Fragestellung zur Überprüfung der Spezifität von DeepGestalt entwickelt und hierfür das Studiendesign einer diagnostischen Genauigkeitsstudie mit negativen (nach Alter, Geschlecht und ethnischem Hintergrund passenden) Kontrollen mitentworfen. Sie hat die Syndrom-Kohorte und die Kontrollkohorte mitaufgebaut (Auswahl der passenden Syndrome und Auswahl passender Bilder) und 135 Proband*innen und die dazugehörigen Kontrollen eingepflegt. Die klinische Phänotypisierung der Kontrollkohorte erfolgte bei diesen durch Frau Hajjir. Weiterhin hat sie das Manuskript mitverfasst.

Pantel, Jean Tori; Hajjir, Nurulhuda; Danyel, Magdalena; Elsner, Jonas; Abad-Perez, Angela Teresa; Hansen, Peter; Mundlos, Stefan; Spielmann, Malte; Horn, Denise; Ott, Claus-Eric; Mensah, Martin Atta,
Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals With and Without a Genetic Syndrome: Diagnostic Accuracy Study, J. Med. Internet Res., 2020

Beitrag im Einzelnen:

- Entwicklung der diagnostischen Genauigkeitsstudie zur Überprüfung der Spezifität von DeepGestalt.
- Aufbau des Studiendesigns.
- Auswahl der Patient*innen und Aufbau der Syndrom-Kohorte.
- Auswahl der unauffälligen Bilder und Aufbau der Kontroll-Kohorte.

- Benennung möglicher Störfaktoren (Geschlecht, Ethnizität) und Exklusion dieser.
- Klinische Interpretation der Daten.
- Mitverfassen des Manuskriptes.

_____

Unterschrift, Datum und Stempel der erstbetreuenden Hochschullehrerin

_____

Unterschrift der Doktorandin

9 Auszug aus der Journal Summary List (ISI Web of Knowledge SM)

Journal Data Filtered By: **Selected JCR Year: 2018** Selected Editions: SCIE,SSCI
Selected Categories: **"MEDICAL INFORMATICS"**
Selected Category Scheme: WoS
Gesamtanzahl: 26 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	JOURNAL OF MEDICAL INTERNET RESEARCH	13,602	4.945	0.030580
2	JMIR mHealth and uHealth	2,576	4.301	0.007920
3	JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	9,319	4.292	0.019480
4	IEEE Journal of Biomedical and Health Informatics	4,082	4.217	0.010320
5	ARTIFICIAL INTELLIGENCE IN MEDICINE	2,462	3.574	0.002960
6	COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE	7,147	3.424	0.009350
7	JMIR Serious Games	269	3.351	0.000660
8	JMIR Medical Informatics	384	3.188	0.001480
9	JOURNAL OF BIOMEDICAL INFORMATICS	7,431	2.950	0.010300
10	MEDICAL DECISION MAKING	5,281	2.793	0.009000
11	INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS	4,765	2.731	0.006720
12	JOURNAL OF MEDICAL SYSTEMS	4,680	2.415	0.006220
13	STATISTICAL METHODS IN MEDICAL RESEARCH	4,156	2.388	0.012230
14	Health Informatics Journal	691	2.297	0.001450
15	BMC Medical Informatics and Decision Making	3,578	2.067	0.008490
16	MEDICAL & BIOLOGICAL ENGINEERING & COMPUTING	5,904	2.039	0.004380
17	STATISTICS IN MEDICINE	24,925	1.847	0.034040

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
18	Health Information Management Journal	320	1.742	0.000390
19	JOURNAL OF EVALUATION IN CLINICAL PRACTICE	4,039	1.536	0.005120
20	INTERNATIONAL JOURNAL OF TECHNOLOGY ASSESSMENT IN HEALTH CARE	2,143	1.418	0.002140
21	Applied Clinical Informatics	664	1.306	0.002050
22	Informatics for Health & Social Care	285	1.218	0.000470
23	CIN-COMPUTERS INFORMATICS NURSING	836	1.029	0.001120
24	METHODS OF INFORMATION IN MEDICINE	1,330	1.024	0.001760
25	Biomedical Engineering-Biomedizinische Technik	1,007	1.007	0.001320
26	Therapeutic Innovation & Regulatory Science	371	0.901	0.001600

10 Druckexemplar der ausgewählten Publikation

10.1 Druckexemplar von *Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals With and Without a Genetic Syndrome: Diagnostic Accuracy Study*

Original Paper

Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals With and Without a Genetic Syndrome: Diagnostic Accuracy Study

Jean Tori Pantel[1,2*]; Nurulhuda Hajjir[1,3*]; Magdalena Danyel[1,4]; Jonas Elsner[1]; Angela Teresa Abad-Perez[1]; Peter Hansen[1,5], PhD; Stefan Mundlos[1,6], Prof Dr; Malte Spielmann[6,7], Prof Dr; Denise Horn[1], Prof Dr; Claus-Eric Ott[1], MD; Martin Atta Mensah[1,8], MD

[1]Institute of Medical Genetics and Human Genetics, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany

[2]Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

[3]Klinik für Pädiatrie mit Schwerpunkt Gastroenterologie, Nephrologie und Stoffwechselmedizin, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany

[4]Berlin Center for Rare Diseases, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany

[5]The Jackson Laboratory for Genomic Medicine, Farmington, CT, United States

[6]RG Development & Disease, Max Planck Institute for Molecular Genetics, Berlin, Germany

[7]Institute of Human Genetics, University of Lübeck, Lübeck, Germany

[8]Berlin Institute of Health, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Martin Atta Mensah, MD
Institute of Medical Genetics and Human Genetics
Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health
Berlin
Germany
Phone: 49 30 450 569 132
Fax: 49 30 450 569 914
Email: martin-atta.mensah@charite.de

Abstract

Background: Collectively, an estimated 5% of the population have a genetic disease. Many of them feature characteristics that can be detected by facial phenotyping. Face2Gene CLINIC is an online app for facial phenotyping of patients with genetic syndromes. DeepGestalt, the neural network driving Face2Gene, automatically prioritizes syndrome suggestions based on ordinary patient photographs, potentially improving the diagnostic process. Hitherto, studies on DeepGestalt's quality highlighted its sensitivity in syndromic patients. However, determining the accuracy of a diagnostic methodology also requires testing of negative controls.

Objective: The aim of this study was to evaluate DeepGestalt's accuracy with photos of individuals with and without a genetic syndrome. Moreover, we aimed to propose a machine learning–based framework for the automated differentiation of DeepGestalt's output on such images.

Methods: Frontal facial images of individuals with a diagnosis of a genetic syndrome (established clinically or molecularly) from a convenience sample were reanalyzed. Each photo was matched by age, sex, and ethnicity to a picture featuring an individual without a genetic syndrome. Absence of a facial gestalt suggestive of a genetic syndrome was determined by physicians working in medical genetics. Photos were selected from online reports or were taken by us for the purpose of this study. Facial phenotype was analyzed by DeepGestalt version 19.1.7, accessed via Face2Gene CLINIC. Furthermore, we designed linear support vector machines (SVMs) using Python 3.7 to automatically differentiate between the 2 classes of photographs based on DeepGestalt's result lists.

Results: We included photos of 323 patients diagnosed with 17 different genetic syndromes and matched those with an equal number of facial images without a genetic syndrome, analyzing a total of 646 pictures. We confirm DeepGestalt's high sensitivity (top 10 sensitivity: 295/323, 91%). DeepGestalt's syndrome suggestions in individuals without a craniofacially dysmorphic syndrome followed a nonrandom distribution. A total of 17 syndromes appeared in the top 30 suggestions of more than 50% of nondysmorphic images. DeepGestalt's top scores differed between the syndromic and control images (area under the receiver operating characteristic [AUROC] curve 0.72, 95% CI 0.68-0.76; $P<.001$). A linear SVM running on DeepGestalt's result vectors showed stronger differences (AUROC 0.89, 95% CI 0.87-0.92; $P<.001$).

Conclusions: DeepGestalt fairly separates images of individuals with and without a genetic syndrome. This separation can be significantly improved by SVMs running on top of DeepGestalt, thus supporting the diagnostic process of patients with a genetic syndrome. Our findings facilitate the critical interpretation of DeepGestalt's results and may help enhance it and similar computer-aided facial phenotyping tools.

(J Med Internet Res 2020;22(10):e19263) doi: 10.2196/19263

KEYWORDS

facial phenotyping; DeepGestalt; facial recognition; Face2Gene; medical genetics; diagnostic accuracy; genetic syndrome; machine learning

Introduction

Background

Although individual genetic diseases are rare, they collectively affect an estimated 5% of a population [1]. Thus, these diseases represent a major challenge for health care systems, as it usually requires highly specialized knowledge to propose a specific genetic diagnosis. Assessing the facial phenotypes of patients with genetic syndromes is key to this diagnostic process [2]. Traditionally performed by a physician, the advents of computer vision and machine learning in medicine enable rapid and automated assessment of a patient's facial traits [3,4]. Numerous facial phenotyping systems have been developed with the potential to aid the diagnostic processes in medical genetics [5-12]. DeepGestalt, the neural network behind Face2Gene CLINIC, which was trained on more than 17,106 images, is thus far the best-investigated and most convenient to use application [11]. Several studies assessed the algorithm's sensitivity, suggesting that it is of a certain quality [11,13-38]. These tests predominantly analyzed images of patients diagnosed with a genetic disorder known to show characteristic facial features. This appears reasonable as DeepGestalt is designed to identify such syndromes. However, it might introduce a bias in conclusions of the system's everyday clinical use since not all individuals seen in a real-life setting belong to the group of patients included in previous studies of DeepGestalt. This may be because (1) the featured syndrome is yet to be analyzed by the system; (2) an individual features a syndrome not associated with a characteristic facies; or (3) an individual has no syndrome at all.

In addition to such evaluations of DeepGestalt's sensitivity, there is a need for studies on its specificity when tested on individuals without craniofacial dysmorphism. As DeepGestalt is not designed to suggest the class label "inconspicuous face" [11], evaluating its clinical specificity is not too trivial a task. Some studies tested the ability of DeepGestalt's methodology to distinguish between facial images with and without a genetic syndrome by constructing user-specific neural networks trained on healthy control images and on images of limited numbers of well-selected genetic disorders using Face2Gene RESEARCH [20,26-28,30,32,34,39-41]. Their results suggested that neural networks such as DeepGestalt may have the potential to differentiate between the 2 classes and may thus be used in diagnosing patients in medical genetics. Such a test could be applied at different stages of the diagnostic process. Patients who want to know if genetic counseling is necessary could use it as a triage test to check whether a suspicion of a genetic disease is justified. Physicians and other medical professionals could similarly use such a test on patients suspected of having a genetic syndrome to narrow down the range of possible diagnoses. Geneticists could use it as an add-on test to further confirm a diagnosis, for example, in the presence of a variant of unknown significance.

Objectives

We aimed to systematically benchmark DeepGestalt's power to discern images of individuals with a dysmorphic genetic syndrome from images of healthy control individuals. For this purpose, we tested the basic prerequisite for the diagnostic usefulness of DeepGestalt, that is, to yield different scores in persons with a conventionally established diagnosis of a genetic syndrome than in persons without a genetic syndrome (H_1: $\mu_{syndromic} \neq \mu_{healthy}$). We also determined DeepGestalt's capacity to distinguish those images by measuring its area under the receiver operating characteristic (AUROC) curve. Furthermore, we aimed to develop and test a machine learning–based approach to improve DeepGestalt's accuracy.

Methods

Selection and Analysis of Portrait Photos

Study Design

To be included in this study, portrait photos had to depict the entire frontal face (from hairline to chin showing both eyes) and no artifact other than glasses. To achieve a vertical positioning of the face, the images were cropped and rotated if necessary. A convenience sample of online accessible images was collected between September 2019 and December 2019, using a methodology adjusted from Ferry et al [8]. Pictures photographed by us were taken at the 2018 meeting of the

Elterninitiative Apertsyndrom und Verwandte Fehlbildungen eV, a parents' initiative on Apert syndrome and related disorders in Germany, after obtaining written informed consents as approved by the ethics committee of the Charité – Universitätsmedizin Berlin (EA2/190/16). Image inclusion was planned before conducting analysis by DeepGestalt. A sample size of the positive and negative class of 105 (N=210) was calculated using G*Power, version 3.1.9.7 (effect size 0.5; □ =.05; power 0.95; allocation ratio 1).

Defining Reference Phenotypes

Only images of individuals reported to be clinically or molecularly diagnosed with a genetic syndrome were labeled as syndromic. When no syndrome was reported and no facial gestalt suggestive of a syndrome was observed, as judged by physicians working in medical genetics, images were labeled as "healthy."

Computer-Aided Facial Phenotyping

Computer-aided facial phenotyping was performed using DeepGestalt version 19.1.7, accessed via Face2Gene CLINIC (FDNA Inc). Neither the class labels nor diagnoses were passed to DeepGestalt. No other phenotypic information but 1 portrait photo per case was entered into the system. DeepGestalt's training set was tested not to contain duplicates of images used in this study, as described previously [42].

Danyel Cohort

The Danyel cohort, originally described by Danyel et al [30], comprises 116 healthy control images.

Syndromic Cohort

This cohort comprises frontal facial images of 17 syndromes. We planned to collect the same number of images for each of these syndromes. A total of 16 of these syndromes were chosen from the 201 distinct suggestions in DeepGestalt's top 30 results lists of the Danyel cohort. Syndromes of different frequencies ranging from 76% (frequently suggested) to 1% (rarely suggested) were selected. In descending order of frequency, these syndromes are as follows: Fragile X syndrome (OMIM: #300624), Angelman syndrome (OMIM: #105830), Rett syndrome (OMIM: #312750), Phelan-McDermid syndrome (OMIM: #606232), Klinefelter syndrome, Beckwith-Wiedemann syndrome (OMIM: #130650), 22q11.2 deletion syndrome (OMIM: #611867), Sotos syndrome (OMIM: #117550), Noonan syndrome (OMIM: PS163950), Loeys-Dietz syndrome (OMIM: PS609192), Williams-Beuren syndrome (OMIM: #194050), Rubinstein-Taybi syndrome (OMIM: PS180849), achondroplasia (OMIM: #100800), Wolf-Hirschhorn syndrome (OMIM: #194190), Pallister-Killian syndrome (OMIM: #601803), and Treacher Collins syndrome (OMIM: PS154500). In addition, we chose Apert syndrome (OMIM: #101200), which was not implied in the Danyel cohort.

Matched Control Cohort

Each photo of the syndromic cohort was matched to an image of an individual without a genetic syndrome by age, sex, and ethnicity to build a cohort of an equal number of control images.

Statistical Evaluation and Classification Experiments

Face2Gene CLINIC returns DeepGestalt's top 30 syndrome suggestions. DeepGestalt associates each suggestion with a Gestalt score [11]. The syndrome suggestions' frequencies, scores, and ranks were statistically evaluated.

Feature Extraction and Vector Construction

All images were labeled by class (syndromic vs healthy). Vectors were built to hold an attribute for any of the syndromes suggested at least once in DeepGestalt's top 30 suggestions. To construct a vector for a given photo, the 30 highest Gestalt scores were assigned to their respective attributes; and the remaining attributes were set to 0 (s. matrix.txt in Multimedia Appendix 1).

Classification

To differentiate between syndromic and healthy portrait photos, we trained linear support vector machines (SVMs) using the LinearSVM class of scikit-learn, version 0.21.3, with default parameters in Python 3.7. To avoid overfitting, training and testing were performed using a leave-1-out classification scheme. Since ethnic background is a possible confounder of DeepGestalt [15,22,26,29,33], we designed classification experiments based on all images, images of White persons, and those of persons with other ethnicities, to benchmark the influence of ethnicity on SVM performance.

To test a possible influence of the number of top ranks considered, classification of all images was run 30 times with the number of considered top Gestalt ranks, ranging from 1 to 30.

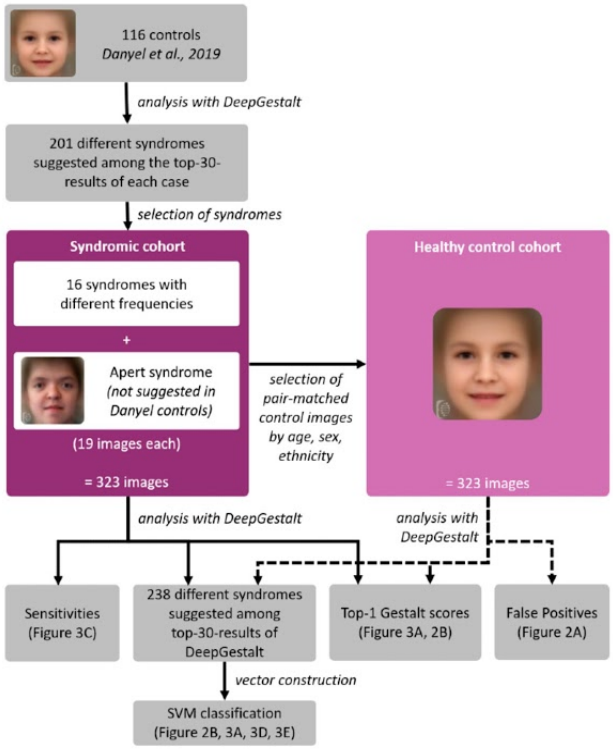
Statistical Analysis

Scores of the syndromic and healthy control cohort were tested to be different using a 2-sided, independent Welch t test. Difference of receiver operating characteristics (ROCs) was tested using a DeLong test. Classification performance was assessed using Matthews correlation coefficient (MCC). All statistical tests were performed in Python 3.7; the code can be found in Multimedia Appendix 1.

Data and Code Availability

The data and code can be found in Multimedia Appendix 1. For reasons of data protection, all data were cumulated (where possible), deidentified, and minimized. Facial images depicted in Figure 1 show computer-generated composite masks and not real individuals. In Multimedia Appendix 1, file data.txt describes the diagnosis, age, sex, and ethnicity of persons in the analyzed set of images; and file matrix.txt contains DeepGestalt's output vectors as used for this study. Files differentiator.py and reproduce.py may be used for reproducing the statistical results of this study. Further information may be found in file readme.txt (Multimedia Appendix 1).

Figure 1. Workflow of classification experiments.



Results

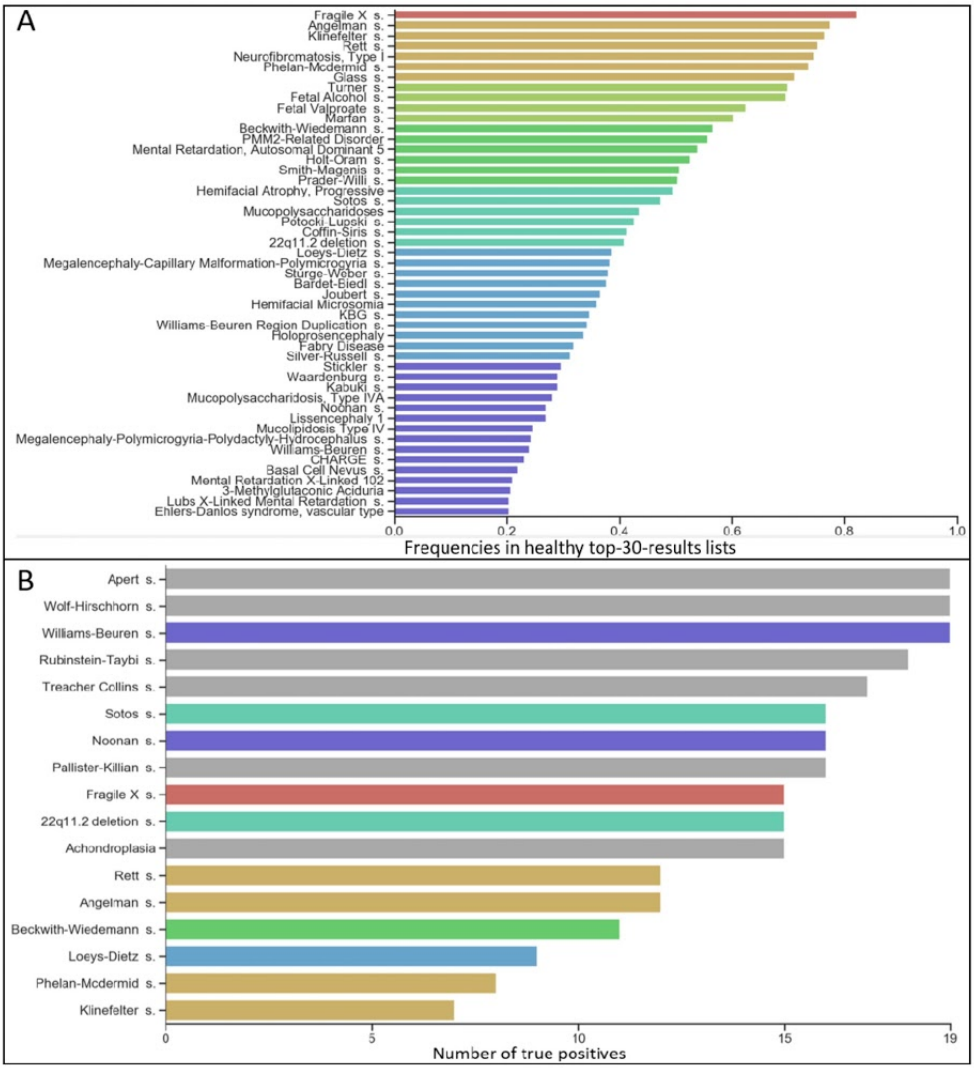
Included Images

We could include 19 images for each of the 17 syndromes in the syndromic cohort. A total of 83% (272/323) of these images were of White persons (file data.txt of Multimedia Appendix 1). Images from the syndromic cohort were matched to 323 images forming the matched control cohort, resulting in a total number of 646 analyzed photos (Figure 1).

Frequencies and Scores of Suggested Syndromes in Control Individuals

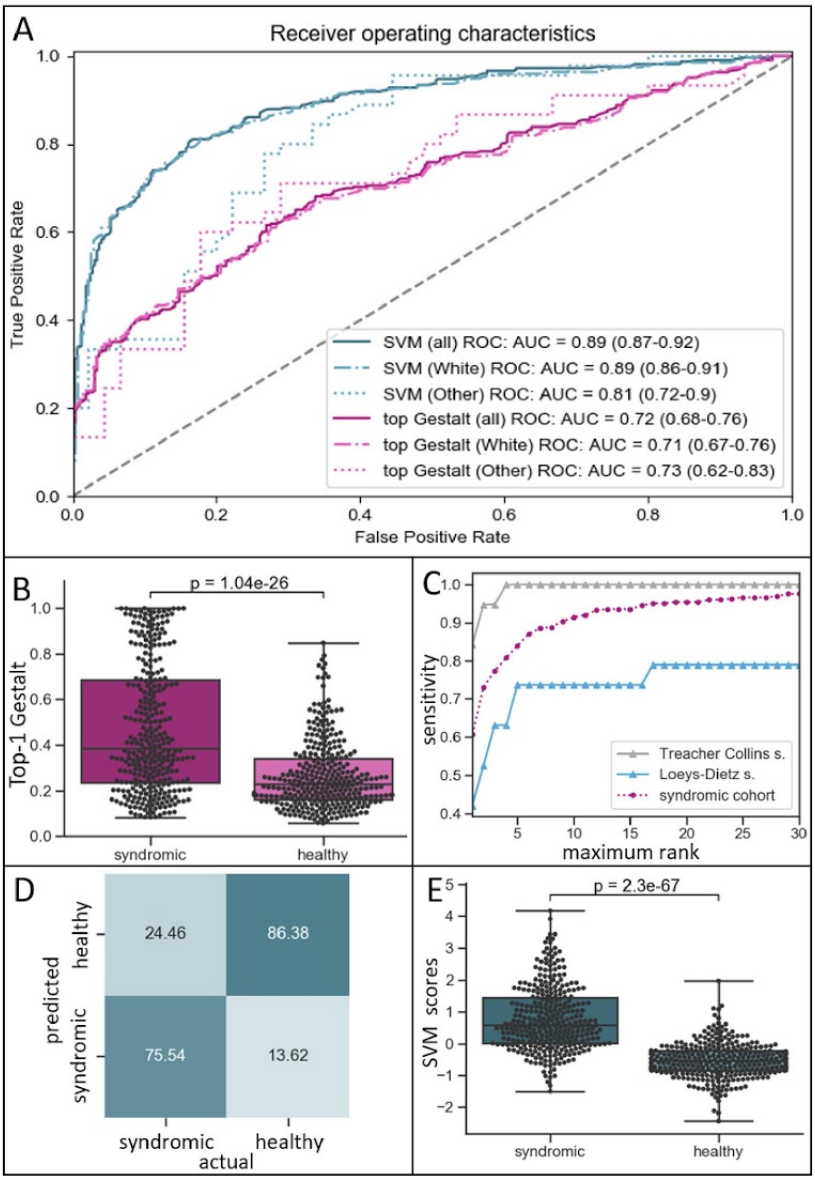
DeepGestalt suggested 238 different syndromes among the top 30 suggestions of the matched control cohort. One syndrome was suggested in more than 80% of the cases (Fragile X syndrome, 82%), 6 syndromes in 70%-80% of the cases; 4 syndromes in 60%-70% of the cases; 6 syndromes in 50%-60% of the cases; 6 syndromes in 40%-50% of the cases; 11 syndromes in 30%-40% of the cases; 15 syndromes in 20%-30% of the cases; 29 syndromes in 10%-20% of the cases; and 160 syndromes at least once in less than 10% of the cases (Figure 2A).

Figure 2. (A) Frequency of syndromes suggested by DeepGestalt in more than 20% of the matched control cohort's top-30-results lists. Colors indicate frequency percentages. (B) Number of images correctly classified as "syndromic"; colors relate to (A) and gray indicates <20%.



The highest first-rank Gestalt score of the matched control cohort amounted to 0.85, and the lowest, to 0.06, with a mean of 0.27 (SD 0.15). First-rank Gestalt scores of the syndromic cohort (highest 1.0; lowest 0.08; mean 0.47, SD 0.28) and the matched control cohort appeared to be separable with an AUROC of 0.72 (95% CI 0.68-0.76) (Figure 3A). Notably, this was found for both tested ethnic groups (Figure 3A, Multimedia Appendix 2), White persons only (AUROC 0.71, 95% CI 0.67-0.76; $P<.001$), and persons of other ethnicities only (AUROC 0.71, 95% CI 0.62-0.83; $P<.001$). Separability of the 2 cohorts is evident and significant ($P<.001$), as shown in Figure 3B.

Figure 3. (A) Receiver operating characteristic (ROC) curves: dashed line indicates random ROC curve; note that support vector machine (SVM) scores yield higher areas under the ROC curves (AUROCs) than their respective raw first-rank Gestalt scores. (B) Distribution of first-rank Gestalt scores in the syndromic cohort and the matched control cohort (healthy). (C) Sensitivities of DeepGestalt (X-axis: number of considered top ranks). Dark-purple circles: average of syndromic cohort; gray triangles: 19 images with Treacher-Collins syndrome; blue triangles: 19 images with Loeys-Dietz syndrome. (D) Distribution of SVM scores in the syndromic cohort and the matched control cohort; note: improved separability as compared to B. (E) SVM classification results based on the entire matched control cohort and syndromic cohort (threshold SVM score: 0).



Sensitivity of DeepGestalt

DeepGestalt's average top 10 sensitivity in the syndromic cohort amounted to 91%, varying between the 17 tested syndromes (Figure 3C, Multimedia Appendix 3). Interestingly, DeepGestalt was sensitive independent of ethnicity (White persons only, 90%; persons of other ethnicities only, 97%). A total of 7 syndromes reached a top 10 sensitivity of 100% (Fragile X, Noonan, Phelan-McDermid, Rett, Sotos, Treacher-Collins, and Williams-Beuren syndromes). DeepGestalt performed worst

for Loeys-Dietz syndrome, with a top 10 sensitivity of 74% (Figure 3C).

Performance of the SVM

Sensitivities of binary SVM classification differed between syndromes (Figure 2B). All images of individuals with Apert syndrome, Wolf-Hirschhorn syndrome, and Williams-Beuren syndrome were correctly classified as being syndromic. The SVM performed worst on the 19 images of individuals with Klinefelter syndrome, correctly classifying only 7 of them as syndromic.

Binary SVM classification of DeepGestalt's output achieved an increased separability of syndromic images and healthy controls as compared to top Gestalt scores with an AUROC of 0.89 (95% CI 0.87-0.92) (Figure 3A). Again, this was true in both tested ethnic groups (Figure 3A), for photos of White persons (AUROC 0.88, 95% CI 0.86-0.91; $P<.001$) and those of persons of other ethnicities (AUROC 0.79, 95% CI 0.62-0.83). However, difference in ROCs was not significant in the latter ($P=.13$). SVM classification performance improved with an increasing number of considered ranks. Using the top 30 Gestalt scores showed the best MCC (0.63), as shown in Multimedia Appendix 4, with a sensitivity of 75.54% and a specificity of 86.38% (Figure 3D). Separability was significant ($P<.001$) (Figure 3E).

Discussion

Classification of Images of Individuals Without a Genetic Syndrome

To our knowledge, this is the first study to systematically analyze DeepGestalt's behavior on portrait photos of individuals without a genetic syndrome. For these images, we show that DeepGestalt's syndrome suggestions follow an interesting distribution. Certain syndromes are implied as differential diagnoses with a considerably high likelihood. Among these were Fragile X, Klinefelter, Rett, and Angelman syndromes, which were suggested in more than 3 quarters of the matched control cohort. In contrast, syndromes such as Treacher-Collins syndrome and Wolf-Hirschhorn syndrome were implied very rarely.

DeepGestalt cannot assign the class label "inconspicuous." Yet, DeepGestalt's scores are used to help judge the presence of a given syndrome. Based on a high maximum Gestalt score, a user could assume that the individual depicted in an entered image is likely to have a syndrome. Likewise, one is tempted to assume that a low maximum Gestalt score makes an underlying syndrome unlikely. Indeed, the mean of first-rank Gestalt scores is higher in images depicting syndromic facies than in images of individuals without a genetic syndrome. Similarly, scores higher than 0.85 appear to be specific indicators of a syndromic facies, and those lower than 0.08 are not suggestive of a genetic syndrome. However, these specific values are very rare. Gestalt scores alone are only fairly sufficient for judging the presence or absence of a genetic syndrome with facial dysmorphism since the distributions of the highest Gestalt scores of the syndromic and matched control cohort greatly overlap. We show that this problem can be reduced by considering both top Gestalt scores and the actual list of suggested syndrome matches. The boost in discriminatory power is illustrated by the increase of the respective AUROCs. Although DeepGestalt cannot directly assess the presence/absence of a syndromic facies, machine learning–based tools (eg, SVMs) built on top of DeepGestalt may be used for this purpose.

It is noteworthy that we achieved promising results with a comparably low number of samples and a low complexity classification model with default hyperparameters. We assume that the quality and complexity of future classifiers will improve as more data will become available. Increasing the number of top ranks considered for vector construction increased the performance of the SVM. However, the number of DeepGestalt's suggestions accessible via Face2Gene CLINIC is limited to 30 suggestions. We hypothesize that using more than just the 30 top ranks for vector construction might further boost classification performance. We classified DeepGestalt's output to predict the presence of a syndromic facies. We also suggest evaluating classification performance based on DeepGestalt's input vectors.

Potential Confounders

Until now, differences in the diagnostic performance of DeepGestalt, which arise due to the ethnicity of the person depicted, have been evaluated using DeepGestalt's sensitivity. Studies of earlier versions of DeepGestalt showed that its sensitivity is dependent on the ethnic background in certain syndromes [15,22]. Studies of more recent versions of DeepGestalt suggested that ethnicity had no major influence on its sensitivity [26,29]. In our set of syndromic images, DeepGestalt's sensitivity is remarkably high, which is in line with the previous studies highlighting DeepGestalt's good general sensitivity [11,36,42]. This high sensitivity of DeepGestalt was confirmed for both groups of images, those of White persons and those of persons of other ethnicities. Improvement of distinguishability of images of individuals with and without a genetic syndrome appeared to be stronger in the group of photos of White persons than in the group of photos of persons of other ethnicities. However, we assume that this is caused by the limited sample size of images of non-White persons in our data set. We believe that our approach is also applicable to populations comprising predominantly other ethnicities.

The SVM had difficulties classifying images of patients with syndromes that were frequently suggested in healthy controls. Possible explanations for DeepGestalt's output to be similar in controls and individuals with these syndromes could be as follows: (1) such syndromes have only mild characteristic facial features; (2) they have a typical facial gestalt, which is present only in some but not all affected individuals; or (3) they have no typical facies at all. For example, not all patients with Loeys-Dietz syndrome exhibit distinctive facial features [43], and the facial appearances of males with Klinefelter syndrome show no commonly observed characteristics [44].

Further Research

Further research is necessary to determine DeepGestalt's capacity to distinguish individuals with and without a genetic syndrome when combined with other sources of information, such as genetic test results and nonfacial phenotypic information. We suggest including additional scores that are based on both phenotype and genotype (eg, prioritization of exome data by image analysis [PEDIA] scores [42]) in future classifiers of the presence/absence of a syndromic facies.

The increasing use and quality of facial phenotyping software in clinical genetics should also be accompanied by an ethical evaluation of these systems [45]. This affects issues such as the automation of medical diagnostic action, the sharing of (potentially identifiable) data, and a potentially altered doctor-patient relationship. In particular, a systematic analysis of the patient perspective on the use of computer-aided facial analysis methodologies in clinical genetics is lacking so far.

We believe that our findings will help improve future versions of DeepGestalt and similar systems and are crucial when interpreting Face2Gene's results in the clinical routine. In particular, we recommend providing users with the false-positive rates of each suggested syndrome.

Conclusion

DeepGestalt is a computer-aided facial phenotyping tool that showed promising results for detecting a potentially syndromic facies. It yields higher first-rank scores in individuals with a genetic syndrome than in those without a diagnosis of a genetic syndrome. Its output may be classified to improve this detection. The exact stage to use DeepGestalt during the diagnostic makeup of individuals with a suspected genetic syndrome remains to be determined. Primarily, it should be used by expert geneticists.

Acknowledgments

We thank the members of the Elterninitiative Apertsyndrom und Verwandte Fehlbildungen eV, a parents' initiative on Apert syndrome and related disorders in Germany, for the contribution of their images, and Yaron Gurovich and Nicole Fleischer of FDNA Inc for technical assistance in checking DeepGestalt's training set for duplicate images used in this study. MAM is a participant in the BIH Charité Digital Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health. We acknowledge support from the German Research Foundation (DFG) and the Open Access Publication Funds of Charité – Universitätsmedizin Berlin.

Authors' Contributions

JTP, NH, and MAM designed the study. JTP, NH, MD, JE, ATAP, and MAM collected the data. SM, MS, DH, and CEO provided insights that were critical for the interpretation of data. MAM implemented the Python code with support from PH. PH and MAM performed the statistical analysis. JTP, NH, CEO, and MAM wrote the manuscript with approval of all the authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Code and data.
[ZIP File (Zip Archive), 137 KB-Multimedia Appendix 1]

Multimedia Appendix 2

(A) Distribution of first-rank Gestalt scores for the images of White persons in the syndromic cohort and the matched control cohort (healthy). (B) Distribution of first-rank Gestalt scores for the images of persons with other ethnicities in the syndromic cohort and the matched control cohort (healthy).
[PNG File , 97 KB-Multimedia Appendix 2]

Multimedia Appendix 3

DeepGestalt's sensitivities: purple circles indicate the average of the entire syndromic cohort; for other symbols/coloring, see respective subfigure title.
[PNG File , 208 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Performance of the SVM on the entire syndromic cohort and matched control cohort: X-axis number of top-rank Gestalt score used for vector construction per case. MCC: Matthews correlation coefficient. Note: rising tendency.
[PNG File , 46 KB-Multimedia Appendix 4]

References

1. Jackson M, Marks L, May GHW, Wilson JB. The genetic basis of disease. Essays Biochem 2018 Dec 03;62(5):643-723 [FREE Full text] [doi: 10.1042/EBC20170053] [Medline: 30509934]
2. Hart TC, Hart PS. Genetic studies of craniofacial anomalies: clinical implications and applications. Orthod Craniofac Res 2009 Aug;12(3):212-220 [FREE Full text] [doi: 10.1111/j.1601-6343.2009.01455.x] [Medline: 19627523]
3. Xie Q, Faust K, Van Ommeren R, Sheikh A, Djuric U, Diamandis P. Deep learning for image analysis: Personalizing medicine closer to the point of care. Crit Rev Clin Lab Sci 2019 Jan;56(1):61-73. [doi: 10.1080/10408363.2018.1536111] [Medline: 30628494]
4. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med 2019 Nov 19;11(1):70 [FREE Full text] [doi: 10.1186/s13073-019-0689-8] [Medline: 31744524]
5. Boehringer S, Vollmar T, Tasse C, Wurtz RP, Gillessen-Kaesbach G, Horsthemke B, et al. Syndrome identification based on 2D analysis software. Eur J Hum Genet 2006 Oct;14(10):1082-1089 [FREE Full text] [doi: 10.1038/sj.ejhg.5201673] [Medline: 16773127]
6. Vollmar T, Maus B, Wurtz RP, Gillessen-Kaesbach G, Horsthemke B, Wieczorek D, et al. Impact of geometry and viewing angle on classification accuracy of 2D based analysis of dysmorphic faces. Eur J Med Genet 2008;51(1):44-53. [doi: 10.1016/j.ejmg.2007.10.002] [Medline: 18054308]
7. Boehringer S, Guenther M, Sinigerova S, Wurtz RP, Horsthemke B, Wieczorek D. Automated syndrome detection in a set of clinical facial photographs. Am J Med Genet A 2011 Sep;155A(9):2161-2169. [doi: 10.1002/ajmg.a.34157] [Medline: 21815261]
8. Ferry Q, Steinberg J, Webber C, FitzPatrick DR, Ponting CP, Zisserman A, et al. Diagnostically relevant facial gestalt information from ordinary photos. Elife 2014 Jun 24;3:e02020 [FREE Full text] [doi: 10.7554/eLife.02020] [Medline: 24963138]
9. Cerrolaza JJ, Porras AR, Mansoor A, Zhao Q, Summar M, Linguraru MG. Identification of dysmorphic syndromes using landmark-specific local texture descriptors Internet. 2016 Presented at: IEEE 13th International Symposium on Biomedical Imaging (ISBI); 13-16 April 2016; Prague, Czech Republic. [doi: 10.1109/isbi.2016.7493453]
10. Tu L, Porras A, Boyle A, Linguraru M. Analysis of 3D Facial Dysmorphology in Genetic Syndromes from Unconstrained 2D Photographs Internet. In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science, vol 11070. Cham: Springer; 2018:347-355.
11. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med 2019 Jan;25(1):60-64. [doi: 10.1038/s41591-018-0279-0] [Medline: 30617323]
12. Dudding-Byth T, Baxter A, Holliday EG, Hackett A, O'Donnell S, White SM, et al. Computer face-matching technology using two-dimensional photographs accurately matches the facial gestalt of unrelated individuals with the same syndromic form of intellectual disability. BMC Biotechnol 2017 Dec 19;17(1):90 [FREE Full text] [doi: 10.1186/s12896-017-0410-1] [Medline: 29258477]
13. Basel-Vanagaite L, Wolf L, Orin M, Larizza L, Gervasini C, Krantz ID, et al. Recognition of the Cornelia de Lange syndrome phenotype with facial dysmorphology novel analysis. Clin Genet 2016 May;89(5):557-563. [doi: 10.1111/cge.12716] [Medline: 26663098]
14. Gripp KW, Baker L, Telegrafi A, Monaghan KG. The role of objective facial analysis using FDNA in making diagnoses following whole exome analysis. Report of two patients with mutations in the BAF complex genes. Am J Med Genet A 2016 Jul;170(7):1754-1762. [doi: 10.1002/ajmg.a.37672] [Medline: 27112773]
15. Lumaka A, Cosemans N, Lulebo Mampasi A, Mubungu G, Mvuama N, Lubala T, et al. Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. Clin Genet 2017 Aug;92(2):166-171. [doi: 10.1111/cge.12948] [Medline: 27925162]
16. Hadj-Rabia S, Schneider H, Navarro E, Klein O, Kirby N, Huttner K, et al. Automatic recognition of the XLHED phenotype from facial images. Am J Med Genet A 2017 Sep;173(9):2408-2414. [doi: 10.1002/ajmg.a.38343] [Medline: 28691769]
17. Gardner OK, Haynes K, Schweitzer D, Johns A, Magee WP, Urata MM, et al. Familial Recurrence of 3MC Syndrome in Consanguineous Families: A Clinical and Molecular Diagnostic Approach With Review of the Literature. Cleft Palate Craniofac J 2017 Nov;54(6):739-748. [doi: 10.1597/15-151] [Medline: 27356087]
18. Valentine M, Bihm DCJ, Wolf L, Hoyme HE, May PA, Buckley D, et al. Computer-Aided Recognition of Facial Attributes for Fetal Alcohol Spectrum Disorders. Pediatrics 2017 Dec;140(6):e20162028. [doi: 10.1542/peds.2016-2028] [Medline: 29187580]
19. Knaus A, Pantel JT, Pendziwiat M, Hajjir N, Zhao M, Hsieh T, et al. Characterization of glycosylphosphatidylinositol biosynthesis defects by clinical features, flow cytometry, and automated image analysis. Genome Med 2018 Jan 09;10(1):3 [FREE Full text] [doi: 10.1186/s13073-017-0510-5] [Medline: 29310717]
20. Liehr T, Acquarola N, Pyle K, St-Pierre S, Rinholm M, Bar O, et al. Next generation phenotyping in Emanuel and Pallister-Killian syndrome using computer-aided facial dysmorphology analysis of 2D photos. Clin Genet 2018 Feb;93(2):378-381. [doi: 10.1111/cge.13087] [Medline: 28661575]

21. Zarate YA, Smith-Hicks CL, Greene C, Abbott M, Siu VM, Calhoun ARUL, et al. Natural history and genotype-phenotype correlations in 72 individuals with SATB2-associated syndrome. Am J Med Genet A 2018 Apr;176(4):925-935. [doi: 10.1002/ajmg.a.38630] [Medline: 29436146]
22. Pantel JT, Zhao M, Mensah MA, Hajjir N, Hsieh T, Hanani Y, et al. Advances in computer-assisted syndrome recognition by the example of inborn errors of metabolism. J Inherit Metab Dis 2018 May;41(3):533-539 [FREE Full text] [doi: 10.1007/s10545-018-0174-3] [Medline: 29623569]
23. Ferreira CR, Altassan R, Marques-Da-Silva D, Francisco R, Jaeken J, Morava E. Recognizable phenotypes in CDG. J Inherit Metab Dis 2018 May;41(3):541-553 [FREE Full text] [doi: 10.1007/s10545-018-0156-5] [Medline: 29654385]
24. Jiang Y, Wangler MF, McGuire AL, Lupski JR, Posey JE, Khayat MM, et al. The phenotypic spectrum of Xia-Gibbs syndrome. Am J Med Genet A 2018 Jun;176(6):1315-1326 [FREE Full text] [doi: 10.1002/ajmg.a.38699] [Medline: 29696776]
25. Graul-Neumann LM, Mensah MA, Klopocki E, Uebe S, Ekici AB, Thiel CT, et al. Biallelic intragenic deletion in MASP1 in an adult female with 3MC syndrome. Eur J Med Genet 2018 Jul;61(7):363-368. [doi: 10.1016/j.ejmg.2018.01.016] [Medline: 29407414]
26. Vorravanpreecha N, Lertboonnum T, Rodjanadit R, Sriplienchan P, Rojnueangnit K. Studying Down syndrome recognition probabilities in Thai children with de-identified computer-aided facial analysis. Am J Med Genet A 2018 Sep;176(9):1935-1940. [doi: 10.1002/ajmg.a.40483] [Medline: 30070762]
27. Martinez-Monseny A, Cuadras D, Bolasell M, Muchart J, Arjona C, Borregan M, et al. From gestalt to gene: early predictive dysmorphic features of PMM2-CDG. J Med Genet 2019 Apr;56(4):236-245. [doi: 10.1136/jmedgenet-2018-105588] [Medline: 30464053]
28. Pascolini G, Fleischer N, Ferraris A, Majore S, Grammatico P. The facial dysmorphology analysis technology in intellectual disability syndromes related to defects in the histones modifiers. J Hum Genet 2019 Aug;64(8):721-728. [doi: 10.1038/s10038-019-0598-0] [Medline: 31086247]
29. Mishima H, Suzuki H, Doi M, Miyazaki M, Watanabe A, Matsumoto T, et al. Evaluation of Face2Gene using facial images of patients with congenital dysmorphic syndromes recruited in Japan. J Hum Genet 2019 Aug;64(8):789-794. [doi: 10.1038/s10038-019-0619-z] [Medline: 31138847]
30. Danyel M, Cheng Z, Jung C, Boschann F, Pantel JT, Hajjir N, et al. Differentiation of MISSLA and Fanconi anaemia by computer-aided image analysis and presentation of two novel MISSLA siblings. Eur J Hum Genet 2019 Dec;27(12):1827-1835. [doi: 10.1038/s41431-019-0469-3] [Medline: 31320746]
31. Pascolini G, Valiante M, Bottillo I, Laino L, Fleischer N, Ferraris A, et al. Striking phenotypic overlap between Nicolaides-Baraitser and Coffin-Siris syndromes in monozygotic twins with ARID1B intragenic deletion. Eur J Med Genet 2020 Mar;63(3):103739. [doi: 10.1016/j.ejmg.2019.103739] [Medline: 31421289]
32. Kruszka P, Hu T, Hong S, Signer R, Cogné B, Isidor B, et al. Phenotype delineation of ZNF462 related syndrome. Am J Med Genet A 2019 Oct;179(10):2075-2082 [FREE Full text] [doi: 10.1002/ajmg.a.61306] [Medline: 31361404]
33. Fung JLF, Rethanavelu K, Luk H, Ho MSP, Lo IFM, Chung BHY. Coffin-Lowry syndrome in Chinese. Am J Med Genet A 2019 Oct;179(10):2043-2048. [doi: 10.1002/ajmg.a.61323] [Medline: 31400053]
34. Weiss K, Lazar HP, Kurolap A, Martinez AF, Paperna T, Cohen L, et al. The CHD4-related syndrome: a comprehensive investigation of the clinical spectrum, genotype-phenotype correlations, and molecular basis. Genet Med 2020 Feb;22(2):389-397. [doi: 10.1038/s41436-019-0612-0] [Medline: 31388190]
35. Zarate YA, Bosanko KA, Gripp KW. Using facial analysis technology in a typical genetic clinic: experience from 30 individuals from a single institution. J Hum Genet 2019 Dec;64(12):1243-1245. [doi: 10.1038/s10038-019-0673-6] [Medline: 31551534]
36. Narayanan DL, Ranganath P, Aggarwal S, Dalal A, Phadke SR, Mandal K. Computer-aided Facial Analysis in Diagnosing Dysmorphic Syndromes in Indian Children. Indian Pediatr 2019 Dec 15;56(12):1017-1019 [FREE Full text] [Medline: 31884430]
37. Latorre-Pellicer A, Ascaso Á, Trujillano L, Gil-Salvador M, Arnedo M, Lucia-Campos C, et al. Evaluating Face2Gene as a Tool to Identify Cornelia de Lange Syndrome by Facial Phenotypes. Int J Mol Sci 2020 Feb 04;21(3):1042 [FREE Full text] [doi: 10.3390/ijms21031042] [Medline: 32033219]
38. Arora V, Puri RD, Bijarnia-Mahay S, Verma IC. Expanding the phenotypic and genotypic spectrum of Wiedemann-Steiner syndrome: First patient from India. Am J Med Genet A 2020 May;182(5):953-956. [doi: 10.1002/ajmg.a.61534] [Medline: 32128942]
39. Carli D, Giorgio E, Pantaleoni F, Bruselles A, Barresi S, Riberi E, et al. NBAS pathogenic variants: Defining the associated clinical and facial phenotype and genotype-phenotype correlations. Hum Mutat 2019 Jun;40(6):721-728. [doi: 10.1002/humu.23734] [Medline: 30825388]
40. Staufner C, Peters B, Wagner M, Alameer S, Barić I, Broué P, et al. Defining clinical subgroups and genotype-phenotype correlations in NBAS-associated disease across 110 patients. Genet Med 2020 Mar;22(3):610-621. [doi: 10.1038/s41436-019-0698-4] [Medline: 31761904]

41. Myers L, Anderlid B, Nordgren A, Lundin K, Kuja-Halkola R, Tammimies K, et al. Clinical versus automated assessments of morphological variants in twins with and without neurodevelopmental disorders. Am J Med Genet A 2020 May 12;182(5):1177-1189. [doi: 10.1002/ajmg.a.61545] [Medline: 32162839]
42. Hsieh T, Mensah MA, Pantel JT, Aguilar D, Bar O, Bayat A, et al. PEDIA: prioritization of exome data by image analysis. Genet Med 2019 Dec;21(12):2807-2814 [FREE Full text] [doi: 10.1038/s41436-019-0566-2] [Medline: 31164752]
43. MacCarrick G, Black JH, Bowdin S, El-Hamamsy I, Frischmeyer-Guerrerio PA, Guerrerio AL, et al. Loeys-Dietz syndrome: a primer for diagnosis and management. Genet Med 2014 Aug;16(8):576-587 [FREE Full text] [doi: 10.1038/gim.2014.11] [Medline: 24577266]
44. Bird RJ, Hurren BJ. Anatomical and clinical aspects of Klinefelter's syndrome. Clin Anat 2016 Jul;29(5):606-619. [doi: 10.1002/ca.22695] [Medline: 26823086]
45. Martinez-Martin N. What Are Important Ethical Implications of Using Facial Recognition Technology in Health Care? AMA J Ethics 2019 Mar 01;21(2):E180-E187 [FREE Full text] [doi: 10.1001/amajethics.2019.180] [Medline: 30794128]

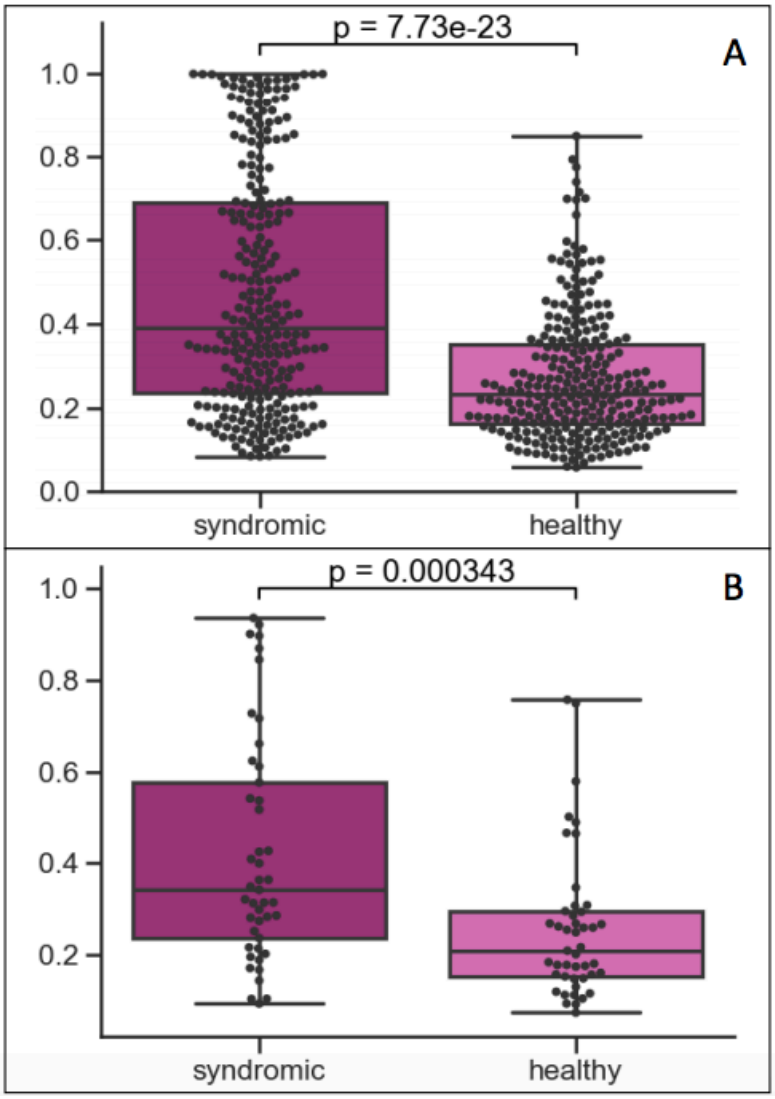
Abbreviations

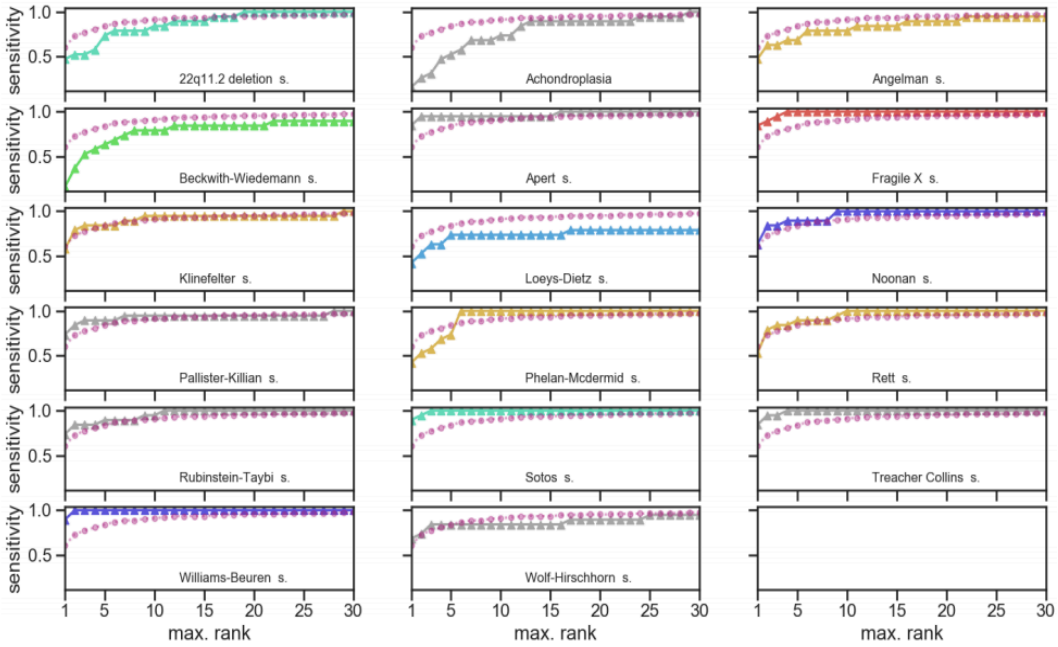
AUROC: area under the receiver operating characteristic
MCC: Matthews correlation coefficient
PEDIA: prioritization of exome data by image analysis
ROC: receiver operating characteristic
SVM: support vector machine

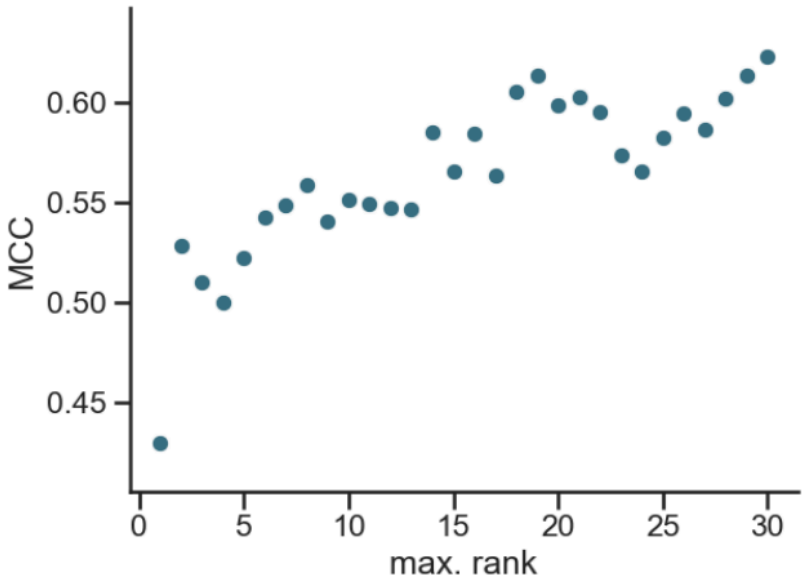
Edited by G Eysenbach; submitted 10.04.20; peer-reviewed by T Liehr, G Pascolini, M Pradhan, D Szinay; comments to author 12.06.20; revised version received 26.06.20; accepted 26.07.20; published 22.10.20

Please cite as:
Pantel JT, Hajjir N, Danyel M, Elsner J, Abad-Perez AT, Hansen P, Mundlos S, Spielmann M, Horn D, Ott CE, Mensah MA
Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals With and Without a Genetic Syndrome: Diagnostic Accuracy Study
J Med Internet Res 2020;22(10):e19263
URL: http://www.jmir.org/2020/10/e19263/
doi: 10.2196/19263
PMID: 33090109

10.2 Anhang zu *Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals With and Without a Genetic Syndrome: Diagnostic Accuracy Study*




sensitivity
1.0
0.5
22q11.2 deletion s.
Achondroplasia
Angelman s.
Beckwith-Wiedemann s.
Apert s.
Fragile X s.
Klinefelter s.
Loeys-Dietz s.
Noonan s.
Pallister-Killian s.
Phelan-Mcdermid s.
Rett s.
Rubinstein-Taybi s.
Sotos s.
Treacher Collins s.
Williams-Beuren s.
Wolf-Hirschhorn s.
1 5 10 15 20 25 30
max. rank


MCC
0.60
0.55
0.50
0.45
0 5 10 15 20 25 30
max. rank

11 Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

12 Publikationsliste

1. Pantel, Jean Tori; Hajjir, Nurulhuda; Danyel, Magdalena; Elsner, Jonas; Abad-Perez, Angela Teresa; Hansen, Peter; Mundlos, Stefan; Spielmann, Malte; Horn, Denise; Ott, Claus-Eric; Mensah, Martin Atta, 2020. "Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals With and Without a Genetic Syndrome: Diagnostic Accuracy Study." *Journal of Medical Internet Research* 22 (10): e19263. (2019 IF: 5.03)

2. Danyel, Magdalena; Cheng, Zhuo; Jung, Christine; Boschann, Felix; Pantel, Jean Tori; Hajjir, Nurulhuda; Flöttmann, Ricarda; Schulz, Solveig; Demuth, Ilja; Sheridan, Eamonn; Mundlos, Stefan; Horn, Denise; Mensah, Martin A, 2019. "Differentiation of MISSLA and Fanconi Anaemia by Computer-Aided Image Analysis and Presentation of Two Novel MISSLA Siblings." *European Journal of Human Genetics: EJHG* 27 (12): 1827–35. (IF 3.657)

3. Hsieh, Tzung-Chien; Mensah, Martin A; Pantel, Jean T; Aguilar, Dione; Bar, Omri; Bayat, Allan; Becerra-Solano, Luis; Bentzen, Heidi B; Biskup, Saskia; Borisov, Oleg; Braaten, Oivind; Ciaccio, Claudia; Coutelier, Marie; Cremer, Kirsten; Danyel, Magdalena; Daschkey, Svenja; Eden, Hilda David; Devriendt, Koenraad; Wilson, Sandra; Douzgou, Sofia; Đukić, Dejan; Ehmke, Nadja; Fauth, Christine; Fischer-Zirnsak, Björn; Fleischer, Nicole; Gabriel, Heinz; Graul-Neumann, Luitgard; Gripp, Karen W; Gurovich, Yaron; Gusina, Asya; Haddad, Nechama; Hajjir, Nurulhuda; Hanani, Yair; Hertzberg, Jakob; Hoertnagel, Konstanze; Howell, Janelle; Ivanovski, Ivan; Kaindl, Angela; Kamphans, Tom; Kamphausen, Susanne; Karimov, Catherine; Kathom, Hadil; Keryan, Anna; Knaus, Alexej; Köhler, Sebastian; Kornak, Uwe; Lavrov, Alexander; Leitheiser, Maximilian; Lyon, Gholson J; Mangold, Elisabeth; Reina, Purificación Marín; Carrascal, Antonio Martinez; Mitter, Diana; Herrador, Laura Morlan; Nadav, Guy; Nöthen, Markus; Orrico, Alfredo; Ott, Claus-Eric; Park, Kristen; Peterlin, Borut; Pölsler, Laura; Raas-Rothschild, Annick; Randolph, Linda; Revencu, Nicole; Fagerberg, Christina Ringmann; Robinson, Peter Nick; Rosnev, Stanislav; Rudnik, Sabine; Rudolf, Gorazd; Schatz, Ulrich; Schossig, Anna; Schubach, Max; Shanoon, Or; Sheridan, Eamonn; Smirin-Yosef, Pola; Spielmann, Malte; Suk, Eun-Kyung; Sznajer,

Yves; Thiel, Christian T; Thiel, Gundula; Verloes, Alain; Vrecar, Irena; Wahl, Dagmar; Weber, Ingrid; Winter, Korina; Wiśniewska, Marzena; Wollnik, Bernd; Yeung, Ming W; Zhao, Max; Zhu, Na; Zschocke, Johannes; Mundlos, Stefan; Horn, Denise; Krawitz, Peter M 2019 "PEDIA: Prioritization of Exome Data by Image Analysis." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 21 (12): 2807–14. (IF 8.904)

4. Pantel, Jean T; Zhao, Max; Mensah, Martin A; Hajjir, Nurulhuda; Hsieh, Tzung-Chien; Hanani, Yair; Fleischer, Nicole; Kamphans, Tom; Mundlos, Stefan; Gurovich, Yaron; Krawitz, Peter M, 2018. "Advances in Computer-Assisted Syndrome Recognition by the Example of Inborn Errors of Metabolism." *Journal of Inherited Metabolic Disease* 41 (3): 533–39. (IF:4.036)

5. Knaus, Alexej; Pantel, Jean Tori; Pendziwiat, Manuela; Hajjir, Nurulhuda; Zhao, Max; Hsieh, Tzung-Chien; Schubach, Max; Gurovich, Yaron; Fleischer, Nicole; Jäger, Marten; Köhler, Sebastian; Muhle, Hiltrud; Korff, Christian; Møller, Rikke S; Bayat, Allan; Calvas, Patrick; Chassaing, Nicolas; Warren, Hannah; Skinner, Steven; Louie, Raymond; Evers, Christina; Bohn, Marc; Christen, Hans-Jürgen; van den Born, Myrthe; Obersztyn, Ewa; Charzewska, Agnieszka; Endziniene, Milda; Kortüm, Fanny; Brown, Natasha; Robinson, Peter N; Schelhaas, Helenius J; Weber, Yvonne; Helbig, Ingo; Mundlos, Stefan; Horn, Denise; Krawitz, Peter M 2018. "Characterization of Glycosylphosphatidylinositol Biosynthesis Defects by Clinical Features, Flow Cytometry, and Automated Image Analysis." *Genome Medicine* 10 (1): 3. (IF: 10.675)

13 Danksagung

Diese Arbeit wäre ohne die Bereitschaft der Patient*innen bzw. ihrer Eltern zur Unterstützung unserer und anderer Forschung nicht möglich gewesen. Ihnen gilt ganz besonderer Dank.

Ich danke Herrn Dr. med. Martin Atta Mensah und Frau Prof. Dr. med. Denise Horn für die exzellente Betreuung und die konstruktive Begleitung meiner Arbeit. Sie standen mir stets unermüdlich mit Rat und Tat zur Seite. Weiterhin danke ich Herrn Prof. Dr. med. Dipl. Phys. Peter M. Krawitz für die Überlassung des Themas und die hervorragende Einarbeitung.

Ich danke außerdem den Co-Autor*innen für ihre unersetzlichen Beiträge und die wunderbaren Impulse in der ganzen Zeit.

Meinen Eltern und meiner Familie kann ich für alles, was sie mir geben, nicht genug danken.