Predictive modelling of facial features from DNA

Master Thesis

G.C.M Jans
MSc Technical Medicine

Universiteit Leiden
TU Delft
Erasmus University Rotterdam

PREDICTIVE MODELLING OF FACIAL FEATURES FROM DNA

Guusje Catherina Maria Jans

Student number: 4318455

September 30th, 2021

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Track: Imaging & Intervention

Department of oral and maxillofacial surgery, Erasmus MC

Department of Radiology and Nuclear Medicine, Erasmus MC

December 2020 – September 2021

Supervisors:

Prof. Dr. Eppo Wolvius

Dr. Gennady Roshchupkin

Thesis committee members:

Prof. Dr. Eppo Wolvius	Erasmus MC	Medical supervisor, Chairman
Dr. Gennady Roshchupkin	Erasmus MC	Technical supervisor
Dr. Xucong Zhang	TU Delft	Independent evaluator

An electronic version of this thesis is available at http://repository.tudelft.nl/

Summary

Background: In recent years, attention to the genetic architecture of normal-range variation in facial morphology has risen and through GWAS genetic loci associated with facial morphology have been identified. However these give no insight in how the face is shaped by genetics. To investigate the relationship between the genotype and the phenotype, predictive modelling can be used. Predictive modelling is a term used to describe genetic prediction models: tools that aim to predict a phenotype from the genotype.

Objectives: The objective of this project was to investigate the possibility of predictive modelling of facial features from DNA. This method could be used to visualize the variation in facial features caused by the underlying genetics

Methods: In this project the LDAK genetic prediction model was used, this is a software package for phenotype prediction, which uses multilinear regression. Two different datasets were used for this study: Generation R and the Rotterdam Study. Generation R data existed of subjects at the age of nine years old and from the Rotterdam study subjects ≥ 45 years old were included. The datasets were processed individually. From both datasets, phenotype and genotype data was used. The phenotype data consisted of 3D facial meshes that were reduced to 200 endophenotypes with an auto-encoder prior to this study, and genotype data consisted of SNPs acquired using genotyping arrays. The prediction model was trained on 90% of the data, the other 10% was used for testing, where the facial morphology was predicted based on the SNPs. To evaluate the prediction a similarity measure was computed between the predicted faces and the ground truth faces. The similarity measure was computed between each predicted face and all ground truth faces in the test set, thereafter they were ranked in ascending order based on the computed similarity. Next the rank of the true ground truth was determined. Based on the ranking, an accuracy plot was constructed for both datasets and the accuracy ratios (AR) were computed.

Results: For the Generation R dataset the AR found was 0.06 for the Generation R dataset and 0.02 for the Rotterdam Study dataset. The results indicate there is some predictive power, however the AR's are only slightly above the lower bound for presence of predictive performance. Furthermore, there was a difference in the AR for the Generation R and Rotterdam Study dataset, which could be the result of increased environmental component in facial morphology which reduced the genetic predictability. However, currently the predictive power is minimal. This could be caused by several factors, such as the number of subjects and prediction model that is restricted to only linear relationships.

Conclusion: The objective of this project was to explore the possibility of predictive modelling of facial features from DNA. It was found that there was some predictive power, however this was very limited. Research on predictive modelling of facial features is still in early stages and further research is required to improve the predictive power.

Table of Contents

PART I
Literature Study

Predictive modelling of facial shape: a literature study

TM30003

Guusje Jans (4318455)
January-February 2021

Supervised by:
Dr. Gennady Roshchupkin
Prof. Dr. Eppo Wolvius

1 Introduction

Like all organisms humans are defined by their phenotype, this encompasses, amongst other things, our physical appearance (1,2). The way we look is part of our identity, it contains information about who we are, for example our age, sex and ancestry. Similarity in physical appearance amongst siblings and relatives, suggest there is hereditary component. Over the years, numerous studies have been conducted investigating the heritability of phenotypic traits (3). The heritability of a trait is determined by the variation in phenotype that can be explained genetically. Family and twin studies are therefore well suited to investigated the genetic component of traits and have been widely used (3–5). The DNA stores the information that is passed on through generations, the inheritable genetic identity (6). In 1990 an international project started to sequence and map the human genome, called the Human Genome Project (7). The Human Genome Project has had an enormous scientific impact and led to more research into human genetic variation and its associations with human health (8). For a long time, the focus of genetic research was on the genetics underlying diseases and anomalies, and less on the genetics of variation in 'normal' phenotypes (9).

The human face is a complex, multipartite trait containing our most visible and distinguished features. It provides the means to communicate with others, show emotions and displays information about our health, age, sex and ancestry. Studies have showed that our facial features are shaped by genetic, epigenetic, environmental factors and their interactions (10–12). Heritability of facial features is, for example, clearly visible in monozygotic twins and, to a lesser extent, amongst related individuals. It has been shown that some parts of the face are more susceptible for environmental factors, while others are predominantly modelled by genetics (10). In recent years, studies have been conducted on the relation between genes and normal-range variation of the face, aiming to develop tools that are able to predict facial morphology based on DNA. This concept is known under different terms in literature: DNA phenotyping, molecular photofitting and predictive modelling. The terms all represent the same concept: predicting the externally visible characteristics of an individually-based on their DNA. The added value of this concept compared to GWAS is the visualisation of variation caused by facial genetics compared to solely the identification of SNPs affecting phenotype without a measure of their effect.

In this literature study I have explored the topic of predictive modelling of facial features based on DNA. I will discuss the potential applications of predictive modelling of the face, the current possibilities of the technique. I will discuss in the genetics of facial morphology and the SNPs that have

been identified. In addition, I will explore the different methodologies to go from genotype to phenotype, what techniques are available and what steps need to be taken.

2 Applications of predictive modelling

Predictive modelling of facial features is still at an early stage; however, it is an interesting subject for multiple fields of study. Development of the concept has clinical, forensic and anthropological relevance. Predictive modelling has numerous ways to add to our medical knowledge and clinical practice. In the past, the focus of genetic research has been on craniofacial anomalies and not on normal-range variation of facial shape and features (13). However, understanding the genetics of normal-range variation of the face is relevant for multiple reasons. First, understanding the genetic architecture of facial features may aide in comprehending the underlying genetic basis of disease traits (14). Second, it could aide in unravelling the contributions of genetic and environmental factors on facial morphology, since both play a role in shaping the facial features (10,11,15). Third, predictive modelling potentially enable identification of shared facial traits, diseases and genes (10,13,16,17). Identifying these facial traits can aide in determining (future) health risks and play a role in diagnostics and making prognoses (13,17).

Predictive modelling of facial features is also an interesting subject for forensic purposes. DNA found at a crime scene could provide investigators with a description of the person of interest when lacking witness reports or when there is no success with conventional DNA profiling, this is called Forensic DNA Phenotyping (FDP) (18–20). FDP could potentially narrow down the list of suspects based on facial appearance (11). To reconstruct a face from a DNA profile comprehension of the genetic architecture and expression mechanisms underlying facial features is required and our knowledge on these subjects is limited, therefore FDP currently remains a long way off (20–23).

Third, predictive modelling of facial features has potential for anthropologic applications (14,18)(18). Ancestry and physical appearance are profoundly related; for example skin colour gives away information on an individual's ancestry (13). Moreover, when comparing individuals from distinct populations differences in facial morphology are clearly observable, however variation is also present within populations (13). Prediction of facial shape of our ancestors could aide in several anthropological topics, e.g. where we come from, how and why we have evolved and maybe even why we look the way we do (4,13,14)

3 Current possibilities of predictive modelling

Numerous studies have been conducted regarding the prediction of externally visible characteristics or EVC's. The current applications of predictive modelling are limited to prediction of a few traits, amongst those are the colour of the eyes, the hair and the skin (21–25). Researchers at the Erasmus University Medical Centre have successfully developed a system that is able to predict these three pigmentation related traits, called the HIrisPlex-S System (26). The system comprises three separate statistical prediction models and the traits are predicted categorically, meaning that for each trait a set of colours is predefined. This varies greatly from the true situation, where eye, hair and skin colour are on a continuous spectrum. Moreover, the model performance varies and performance best for blue and brown eye colour, red hair and dark-to-black skin. Table 1 provides an overview of the EVC's that are possible to predict based on DNA.

Advances have been made in the prediction of other EVC's such as baldness/hair loss, hair structure, presence of moles, height and age, but these attempts have not been successful yet (11,22,30). Scientific research into these traits is at different stages, for a few traits research is focused on finding

the underlying genetic architecture, while for other traits prediction models have been developed but are not distinctive (22).

Publication	Externally Visible Characteristic	No. of SNPs	Specification
Keating et al., 2013	Sex	N/A	Determine male/female
	Biographic ancestry	N/A	Categorical prediction, 5 categories (Africa, Europe, S. Asia, E. Asia, S. America.)
Chaitanya et al., 2018	Eye colour	*24*	Categorical prediction, 3 eye colour categories
	Hair colour	*24*	Categorical prediction, 4 hair colour categories
	Skin colour	*36*	Categorical prediction, 5 skin colour categories
Kukla-Bartoszek et al., 2015	Freckling	*10*	Categorical prediction, non-, medium- or heavily-freckled
Liu et al., 2014	Height	*180*	Predict tall stature (AUC = 0,75)

Table 1. This table provides an overview of the EVC's that can be predicted based on DNA (26–29).

4 Facial Genetics

The embryological development of the external face starts during the 4th gestational week, when six pharyngeal arches arise, and the development is closely related to neural crest cells (13,31–33). These are pluripotent cells that are major contributors to the facial tissues, they will differentiate into bone, cartilages and the majority of the facial connective tissues (34,35). Between the 4th and 8th week major development takes place through a series of highly coordinated and precisely timed events (13,32,36). By week 8 the externally visible facial features have formed and the skeleton of the face is cartilaginous and by week 12 ossification of the face and cranium is in an advanced stage (37,38). Post-natally, the face develops following general somatic growth, with periods of steady growth mixed with periods of rapid growth (13). The embryological development of the face is well known, meanwhile the underlying genetics of facial features remain largely unknown.

Understanding the genetic architecture of facial morphology could help us understand the origin of normal-range variation, differences within and between populations and possibly uncovering the genes underlying medical conditions. There is substantial evidence that there is a genetic component in the shape of facial features, for example differences between males and females or resemblances amongst (monozygotic) twins and between relatives (4,11,13,39). As discussed in section two understanding the genetic architecture of the face is relevant for different reasons. In recent years, much effort has gone into finding the genetics underlying facial features (13). There are different types of studies have contributed to the identification of genes affecting facial morphology (4). Animal, dysmorphology, population and familial studies can uncover genetic loci underlying facial morphology (4). Once potential marker genes have been identified, there association with the face can be further investigated. Finding associations between genes and traits can be done using Genome-Wide Association Studies (GWAS) (4). A GWAS tests single-nucleotide polymorphisms (SNPs) for associations with phenotypes. As a result of eleven GWAS 271 SNPs affecting facial morphology have been identified so far (4,13,33,40–49). Figure 1 provides an overview of the number of SNPs found over the last years. As visible in the graph, the number of associated SNPs has progressed slowly up to this January, when a study identified 203 SNPs associated to facial morphology (48). While various

populations have been included in the GWAS the studied population was predominantly European, with seven studies including solely subjects of European ancestry. Studies including different populations have shown that allele frequencies at identified SNPs have large differences between populations (44,50). The distinctive differences in facial feature between populations may be reflect the underlying genetic variation (51).

When looking at the associations between SNPs and facial regions, it stands out that multiple SNPs are associated to the same facial region, e.g. the nose bridge or chin protrusion. This highlights the polygenic nature of the face, where traits are influenced by different, interacting genes (11). The contributions of the each SNP to the shared traits are unknown and in prediction models it is often assumed that all contribute equally (52). However, it is more likely that the contributions aren't equally distributed, but vary amongst the SNPs (13).

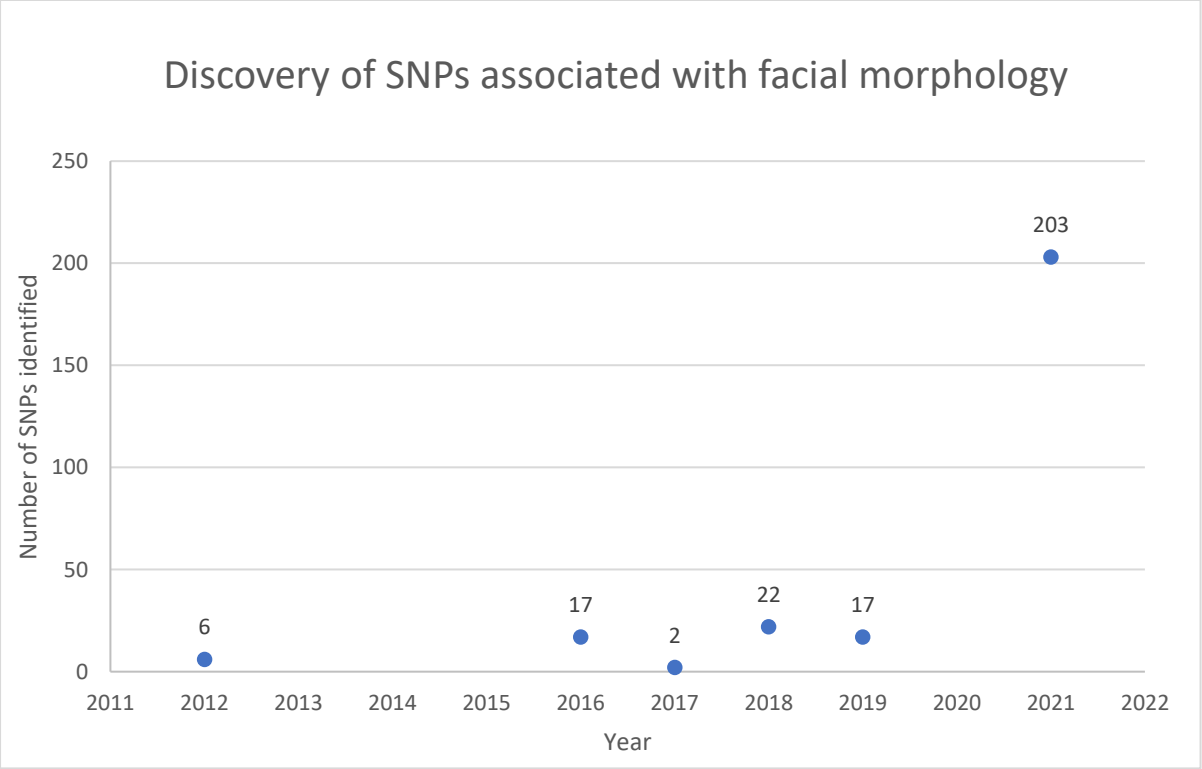

Figure 1. Overview of the progress in the number of SNPs associated to facial morphology over the years.

5 From genotype to phenotype

5.1 Quantification of facial phenotype

To predict the facial phenotype it is necessary to measure facial traits by quantitative means. There are several methods for facial quantification, however due to the complex morphology of the face this is challenging task. The accuracy and reproducibility of the quantification is crucial to ensure meaningful results in studies regarding genetics and facial phenotype. One method to quantify facial phenotype is landmarking. Landmarking is the process of identifying locations on a face that can be reproduced on a second representation, for example the inner corner of the eye. Landmarking can be performed on both 2D and 3D facial data (53). Two types of landmarks exist: anatomical landmarks

and pseudo-anatomical landmarks. The first ones are defined anatomically, while the second type is computed through mathematical definitions. The procedure of landmarking can be either manual or automated (53). Manual annotation is sensitive to inter- and intra-user variations and error-prone (53–55). Therefore, automatic methods are preferred and different (semi)automated methods have been developed over the years (53,54). However, human influence cannot be fully eradicated, as many of the automatic methods are trained on manually annotated faces.. Böhringer et al. has grouped the different methods into four classes: template based, active shape, deep learning and generative (53). Each of the classes with their own advantages and disadvantages. An overview of the four classes can be found in table 2. Even though landmarking has improved with the development of automatic methods, a number of issues remain and landmarking is therefore not the most desirable method to quantify facial phenotype. First, landmarks are insufficient to describe curves and surfaces such as the cheekbones and forehead, since landmark positions along such shapes are not homologous and therefor reproduction across individuals is impossible (56,57). Second, landmarking focusses on a local level of facial phenotype and does not quantify the larger, more general aspects, which are believed to be of great importance to facial appearance (53,58).

Landmarking method	Specification	Strengths	Weaknesses
Template based models	These models use a template for a landmark that is derived from training samples and will be used for comparison on a target image. A template is can be average image patches around landmarks or a full set of image patches. The template can be raw image pathes or a transformation like a wavelet transform.	Can handle texture information very well	This methods is very dependent on the image registration. Geometrically not flexible: poor performance when the image differs greatly from training data.
Active shape models	Active shape models place a graph on an image and try to align it with the existing edges. The graph is often trained on a sample set, in this way it has seen 'possible shapes'. This improves the procedure	Has the ability to capture and predict facial expression	Performance depends on distinctive and well-recognizable edges
Deep learning	Deep learning models consists of algorithms that iteratively estimate parameters in order to optimize the outcome of the learning objective.	Pre-processing of the input images is not required.	The models require large trainingsets, need careful tuning and the algortihm is a 'black box'.
Generative models	Alternative approaches that do not belong to the other categories. For example atlas based methods, where an atlas containing the	NA	NA

Table 1. An overview of the different landmarking methods described by Böhringer et al. (2019), a specification of the methods, their strengths and weaknesses (53).

A different method to quantify facial phenotype are dense surface models (DSM's). A DSM of a face is a digital representation of the face consisting of points in a 3D-space (59). There are different automatic methods developed to create DSM's of faces. One of the methods to build a DSM of a face is described by Hutton et al. (60). They start with nine manually placed landmarks which are biologically homologous, such as the inner and outer corners of the eyes and mouth. From there they build a 3D point distribution model using thin-plate spline warping. In figure 2 a dense surface model is shown created with Hutton's method (60).

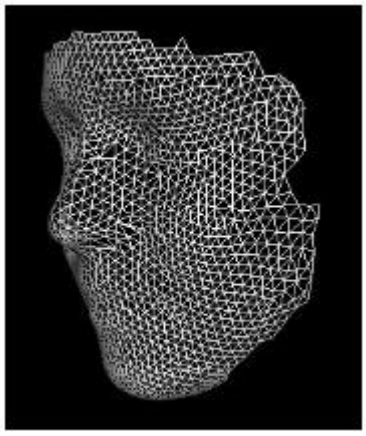

Figure 2. A dense surface model created by Hutton et al. (60)

Despite of the progress in methods for quantification of facial phenotype, it remains a complex task. One of the difficulties of quantification of facial shape is the influence of pose and emotions as the both lead to variations in the face (61).

5.2 Prediction models

There are several approaches to predicting traits from SNPs, in this section I will discuss the different models proposed in literature and the necessity of representative heritability models. Most of the prediction models discussed in literature are an extension to linear regression. Zhang et al. (2020) tested four different prediction models for 14 different traits, binary, categorical as well as continuous traits (52). They developed software containing the prediction models: lasso regression, ridge regression, Bolt-LMM (linear mixed model) and BayesR. These are all extensions to linear regression, and each model has different prior distributions regarding the SNP effect sizes. There was no tool with the best overall performance, the best-performing tool depended on the trait predicted. However, in all cases it was either Bolt-LMM or BayesR. The Bolt-LMM uses a Bayesian linear mixed model to test the association between SNPs and traits. The linear mixed model contains both the fixed and random effects attributed to SNPs (62). BayesR or Bayesian linear regression uses probability distributions rather than point estimates, it assumes the outcome comes from a probability distribution (63). The estimated outcome is improved as more data is gathered. The method of BayesR was also implemented in a prediction model by Lloyed-Jones et al. (64). In these two studies they did not aim to predict facial characteristics, therefore it is unknown if the used prediction models are fitting for facial phenotype. However, they did include a number of the phenotypes that were continuous as is the facial phenotype. The highest reported R^2 values for the continuous traits by the Bolt-LMM and BayesR prediction models were 0.345 and 0.352 respectively. Thus the effect size of the prediction models is weak, however this can be caused by varies reasons and does not have to be the direct effect of the algorithms.

Claes (2014) and Fagertun (2015) did compute prediction models for facial phenotypes (11,12). Claes et al. (2014) used an extension to linear regression: bootstrapped response-based imputation modelling (BRIM) (12). The algorithm uses response variables to compute one or more predictor variables: the response-based imputed predictor variable (RIP). The RIP variables models the effect of input variables on facial phenotype, with input variables being sex for example. A principal component analysis was performed to construct a model of the face from a 3D representation of the face, for each PCA a RIP variable was computed. They found that ancestry and sex had a higher effect size than the individual SNPs. This directly issues the shortcomings of the method; they don't account for polygenic effects. By testing each SNP individually the effect of gene interactions is not taken into account, whilst

it is known that the face has of a the polygenic nature (48). Fagertun et al. (2015) applied linear regression in their prediction model (11). First they computed 37 principal components of the facial phenotype, the shape components. They performed their research with 2D data. A GWAS was performed on the shape components and determined the genetic principal components. Thereafter, the performed linear regression on each shape component. The found 6 shape components that could be predicted with statistical significance. The complexity of the face is not well represented by 2D data, which makes it difficult to determine the suitability of this method of predictive modelling of facial phenotype.

Beside the prediction algorithm, there are other factors that contribute to the performance of the prediction. First, the heritability model used in the prediction model. The heritability model describes the expected heritability contributed by each SNP. It is often assumed that all associated SNPs have the same contribution (52). Zhang et al. (2020) has proven that by using different heritability models, for example heritability models where the contribution of a SNP is attributed to minor allele frequencies or local levels of linkage disequilibrium and functional annotations, the performance of the prediction models increases (52). Thus, taking the polygenic nature of traits into account improves the performance of the prediction model. Second, the sample size for training the prediction models influences the performance of prediction models. Zhang et al (2020) tested the effect of enlarging the sample size and found that enlarging the training set by a quarter led to an 19% increase in the effect size (R^2) of the prediction model (52).

6 Discussion

Predictive modelling of facial phenotype from DNA is a rapidly evolving field, which has gotten more and more attention over the past years. There are several reasons as to why it is an interesting subject, as it has potential applications for multiple fields. First, predictive modelling is an interesting subject for the medical field and there are several potential applications for the technique within this field. Second, the technique is very interesting for forensic purposes as it could aide in reconstructing the appearance of a suspect. Third, predictive modelling could be used for several anthropological applications, such as how and why humans have evolved. Currently, there are a few prediction tools available for externally visible characteristics based on genetic information, such as the colour of the eyes, the hair and the skin. Researchers have tried to predict other traits, such baldness pattern and hair structure, however attempts have been futile so far. To predict facial phenotype, understanding of the underlying genetic architecture is required. In 2012 the first SNPs associated with facial regions were identified, two separate studies found six SNPs in total. Currently there have been 271 SNPs identified, with some SNPs overlapping the associated regions. This highlights the polygenic nature of the face, where several genes affect a feature. In addition to understanding the genetic architecture of the face, two other components are crucial for predictive modelling: a quantification method of facial phenotype and a prediction algorithm. First, The face is a complex structure which makes it difficult to capture the shape correctly, one of the methods to quantify the facial phenotype is landmarking. However, landmarks are suitable to capture curves and surfaces, and are therefor not preferable for predictive modelling. A different method is dense surface models, which are a representation of the face in a 3D space and is better suited for capturing the complex morphology. Second, the prediction method has to model the relationship between the SNPs and the phenotype. The prediction methods currently used are extensions to linear regression, the performance of the methods remains low and the models have a small effect size.

Due to the complexity of the genetic architecture, complex morphology of the face and the unknown relationship of between the genotype and phenotype, predictive modelling is very challenging task.

The prediction methods currently used have not reached high effect size. Artificial intelligence, and deep learning in particular, can be a future direction for predictive modelling of the face. Deep learning is very well suited for challenging problems because of its ability to handle large datasets and model complex relationships. Novel deep learning algorithms allow for interpretable neural networks and enable the linkage of genes to specific traits. Therefor, exploring these techniques could help in the further development of predictive modelling of facial shape from DNA.

7 References

1. Austin christopher. Phenotype [Internet]. Available from: https://www.genome.gov/genetics-glossary/Phenotype

2. Churchill FB. William Johannsen and the genotype concept. In: Journal of the History of Biology. Kluwer Academic Publishers; 1974. p. 5–30.

3. Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond [Internet]. Vol. 3, Nature Reviews Genetics. Nat Rev Genet; 2002. p. 872–82. Available from: https://pubmed.ncbi.nlm.nih.gov/12415317/

4. Roosenboom J, Hens G, Mattern BC, Shriver MD, Claes P. Exploring the Underlying Genetics of Craniofacial Morphology through Various Sources of Knowledge. Biomed Res Int [Internet]. 2016;2016:3054578. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28053980

5. Heritability (Stanford Encyclopedia of Philosophy) [Internet]. Available from: https://plato.stanford.edu/entries/heredity/

6. What is genotype? What is phenotype? – pgEd [Internet]. Available from: https://pged.org/what-is-genotype-what-is-phenotype/

7. What is the Human Genome Project? [Internet]. Available from: https://www.genome.gov/human-genome-project/What

8. Belmont JW, Boudreau A, Leal SM, Hardenbol P, Pasternak S, Wheeler DA, et al. A haplotype map of the human genome. Nature. 2005 Oct 27;437(7063):1299–320.

9. Hernandez LM, Blazer DG, Institute of Medicine (US) Committee on Assessing Interactions Among Social B and GF in H. Genetics and Health. 2006. Available from: https://www.ncbi.nlm.nih.gov/books/NBK19932/

10. Djordjevic J, Zhurov AI, Richmond S. Genetic and environmental contributions to facial morphological variation: A 3D population-based twin study. PLoS One. 2016;

11. Fagertun J, Wolffhechel K, Pers TH, Nielsen HB, Gudbjartsson D, Stefansson H, et al. Predicting facial characteristics from complex polygenic variations. Forensic Sci Int Genet. 2015 Nov 9;19:263–8.

12. Claes P, Liberton DK, Daniels K, Rosana KM, Quillen EE, Pearson LN, et al. Modeling 3D Facial Shape from DNA. Luquetti D, editor. PLoS Genet [Internet]. 2014 Mar 20;10(3):e1004224. Available from: https://dx.plos.org/10.1371/journal.pgen.1004224

13. Richmond S, Howe LJ, Lewis S, Stergiakouli E, Zhurov A. Facial genetics: A brief overview [Internet]. Vol. 9, Frontiers in Genetics. Frontiers Media S.A.; 2018. p. 462. Available from: www.frontiersin.org

14. Walsh S. DNA Phenotyping: The Prediction of Human Pigmentation Traits from Genetic Data [Internet]. 2013. Available from: http://www.rpi.edu/dept/NewsComm/sub/photos/dna_eye.jpg

15. Sero D, Zaidi A, Li J, White JD, Zarzar TBG, Marazita ML, et al. Facial recognition from DNA using face-to-DNA classifiers. Nat Commun [Internet]. 2019 Dec 11;10(1):2557. Available from: http://www.nature.com/articles/s41467-019-10617-y

16. Stephen ID, Hiew V, Coetzee V, Tiddeman BP, Perrett DI. Facial Shape Analysis Identifies Valid Cues to Aspects of Physiological Health in Caucasian, Asian, and African Populations. Front Psychol [Internet]. 2017 Oct 30 ;8(OCT):1883. Available from: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.01883/full

17. Nunes LA, De Jesus AS, Casotti CA, De Araújo ED. Geometric morphometrics and face shape characteristics associated with chronic disease in the elderly. Biosci J. 2018 Mar 1;34(2):1035–46.

18. Walsh S, Lindenbergh A, Zuniga SB, Sijen T, De Knijff P, Kayser M, et al. Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence. Forensic Sci Int Genet. 2011 Nov 1;5(5):464–71.

19. Canales Serrano A. Forensic DNA phenotyping: A promising tool to aid forensic investigation. Current situation. Spanish J Leg Med. 2020 Oct 1;46(4):183–90.

20. Walsh SJ. Recent advances in forensic genetics. In: Expert Review of Molecular Diagnostics [Internet]. Taylor & Francis; 2004. p. 31–40. Available from: https://www.tandfonline.com/action/journalInformation?journalCode=iero20

21. Marano LA, Fridman C. DNA phenotyping: current application in forensic science. Res Reports Forensic Med Sci [Internet]. 2019 Feb 8;Volume 9:1–8. Available from: www.dovepress.com

22. Kayser M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Sci Int Genet. 2015 Feb 11;18:33–48.

23. Stephan CN, Caple JM, Guyomarc'h P, Claes P. An overview of the latest developments in facial imaging [Internet]. Vol. 4, Forensic Sciences Research. Taylor and Francis Ltd.; 2019. p. 10–28. Available from: https://www.tandfonline.com/action/journalInformation?journalCode=tfsr20

24. Schneider PM, Prainsack B, Kayser M. The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry [Internet]. Vol. 116, Deutsches Arzteblatt International. Deutscher Arzte-Verlag GmbH; 2019. p. 873–80. Available from: /pmc/articles/PMC6976916/?report=abstract

25. Curtis C, Hereward J. How Accurately Can Scientists Reconstruct A Person's Face From DNA? | Innovation | Smithsonian Magazine [Internet]. 2018. Available from: https://www.smithsonianmag.com/innovation/how-accurately-can-scientists-reconstruct-persons-face-from-dna-180968951/

26. Chaitanya L, Breslin K, Zuñiga S, Wirken L, Pośpiech E, Kukla-Bartoszek M, et al. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. Forensic Sci Int Genet. 2018 Jul 1;35:123–35.

27. Kukla-Bartoszek M, Pośpiech E, Woźniak A, Boroń M, Karłowska-Pik J, Teisseyre P, et al. DNA-based predictive models for the presence of freckles. Forensic Sci Int Genet. 2019 Sep 1;42:252–9.

28. Liu F, Hendriks AEJ, Ralf A, Boot AM, Benyi E, Sävendahl L, et al. Common DNA variants predict tall stature in Europeans. Hum Genet [Internet]. 2014 Nov 20;133(5):587–97. Available from: http://www.tno.nl/groei

29. Keating B, Bansal AT, Walsh S, Millman J, Newman J, Kidd K, et al. First all-in-one diagnostic tool for DNA intelligence: Genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip. Int J Legal Med [Internet]. 2013 May 13;127(3):559–72. Available from: https://link.springer.com/article/10.1007/s00414-012-0788-1

30. Li X, Li W, Xu Y. Human age prediction based on DNA methylation using a gradient boosting regressor. Genes (Basel) [Internet]. 2018 Sep 1;9(9). Available from: /pmc/articles/PMC6162650/

31. Barnes S. Development of the Face - Nose - Palate - Cleft Lip - TeachMeAnatomy [Internet]. 2020. Available from: https://teachmeanatomy.info/the-basics/embryology/head-neck/face-palate/

32. Som PM, Naidich TP. Illustrated review of the embryology and development of the facial region, part 1: Early face and lateral nasal cavities [Internet]. Vol. 34, American Journal of Neuroradiology. AJNR Am J Neuroradiol; 2013. p. 2233–40. Available from: https://pubmed.ncbi.nlm.nih.gov/23493891/

33. Claes P, Roosenboom J, White JD, Swigut T, Sero D, Li J, et al. Genome-wide mapping of global-to-local genetic effects on human facial shape. Nat Genet [Internet]. 2018 Mar 1;50(3):414–23. Available from: https://doi.org/10.1038/s41588-018-0057-4

34. Cordero DR, Brugmann S, Chu Y, Bajpai R, Jame M, Helms JA. Cranial neural crest cells on the move: Their roles in craniofacial development. Am J Med Genet Part A [Internet]. 2011 Feb;155(2):270–9. Available from: https://pubmed.ncbi.nlm.nih.gov/21271641/

35. (No Title) [Internet]. Available from: https://dev.biologists.org/content/develop/135/23/e2306.full.pdf

36. Ansari A, Bordoni B. Embryology, Face [Internet]. StatPearls. StatPearls Publishing; 2020. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/31424786

37. Som PM, Naidich TP. Illustrated Review of the Embryology and Development of the Facial Region, Part 2: Late Development of the Fetal Face and Changes in the Face from the Newborn to Adulthood. Available from: http://dx.doi.org/10.3174/ajnr.A3414

38. Martini F. Martini's Atlas of the human body. 9th ed. Pearson; 2012.

39. Weinberg SM, Parsons TE, Marazita ML, Maher BS. Heritability of face shape in twins: a preliminary study using 3D stereophotogrammetry and geometric morphometrics. Dent 3000 [Internet]. 2013 Nov 25;1(1):7–11. Available from: /pmc/articles/PMC3911821/?report=abstract

40. Paternoster L, Zhurov AI, Toma AM, Kemp JP, St. Pourcain B, Timpson NJ, et al. Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. Am J Hum Genet [Internet]. 2012 Mar 9; 90(3):478–85. Available from: https://pubmed.ncbi.nlm.nih.gov/22341974/

41. Liu F, van der Lijn F, Schurmann C, Zhu G, Chakravarty MM, Hysi PG, et al. A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. Gibson G, editor. PLoS Genet [Internet]. 2012 Sep 13 ;8(9):e1002932. Available from: https://dx.plos.org/10.1371/journal.pgen.1002932

42. Cole JB, Spritz RA. The Genetics of Facial Morphology. eLS [Internet]. 2017 Sep 15;1–9. Available from: http://doi.wiley.com/10.1002/9780470015902.a0027240

43. Shaffer JR, Orlova E, Lee MK, Leslie EJ, Raffensperger ZD, Heike CL, et al. Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. Barsh GS, editor. PLOS Genet [Internet]. 2016 Aug 25;12(8):e1006149. Available from: https://dx.plos.org/10.1371/journal.pgen.1006149

44. Cha S, Lim JE, Park AY, Do JH, Lee SW, Shin C, et al. Identification of five novel genetic loci related to facial morphology by genome-wide association studies. BMC Genomics [Internet]. 2018 Jun 19;19(1):481. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4865-9

45. Crouch DJM, Winney B, Koppen WP, Christmas WJ, Hutnik K, Day T, et al. Genetics of the human face: Identification of large-effect single gene variants. Proc Natl Acad Sci U S A [Internet]. 2018 Jan 23;115(4):E676–85. Available from: https://pubmed.ncbi.nlm.nih.gov/29301965/

46. Lee MK, Shaffer JR, Leslie EJ, Orlova E, Carlson JC, Feingold E, et al. Genome-wide association study of facial morphology reveals novel associations with FREM1 and PARK2. Li Y, editor. PLoS One [Internet]. 2017 Apr 25;12(4):e0176566. Available from: https://dx.plos.org/10.1371/journal.pone.0176566

47. Xiong Z, Dankova G, Howe LJ, Lee MK, Hysi PG, De Jong MA, et al. Novel genetic loci affecting facial shape variation in humans. Elife. 2019 Nov 1;8.

48. White JD, Indencleef K, Naqvi S, Eller RJ, Hoskens H, Roosenboom J, et al. Insights into the genetic architecture of the human face. Nat Genet [Internet]. 2021 Jan 1;53(1):45–53. Available from: https://doi.org/10.1038/s41588-020-00741-7

49. Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacón-Duque JC, et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nat Commun [Internet]. 2016 Mar 1;7(1):1–12. Available from: www.nature.com/naturecommunications

50. Adhikari K, Fuentes-Guajardo M, Quinto-Sánchez M, Mendoza-Revilla J, Camilo Chacón-Duque J, Acuña-Alonzo V, et al. A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. Nat Commun [Internet]. 2016 May 19;7(1):1–11. Available from: www.nature.com/naturecommunications

51. Hopman SMJ, Merks JHM, Suttie M, Hennekam RCM, Hammond P. Face shape differs in phylogenetically related populations. Eur J Hum Genet [Internet]. 2014 Nov 5;22(11):1268–71. Available from:

https://pubmed.ncbi.nlm.nih.gov/24398794/

52. Zhang Q, Privé F, Vilhjálmsson B, Speed D. Improved genetic prediction of complex traits from individual-level data or summary statistics. 2020; Available from: https://doi.org/10.1101/2020.08.24.265280

53. Böhringer S, De Jong MA. Quantification of facial traits. Frontiers in Genetics. 2019.

54. Tong Y, Liu X, Wheeler FW, Tu PH. Semi-supervised facial landmark annotation. Comput Vis Image Underst. 2012 Aug 1;116(8):922–35.

55. White JD, Ortega-Castrillón A, Matthews H, Zaidi AA, Ekrami O, Snyders J, et al. MeshMonk: Open-source large-scale intensive 3D phenotyping. Sci Rep [Internet]. 2019 Dec 1;9(1). Available from: https://pubmed.ncbi.nlm.nih.gov/30988365/

56. Gunz P, Mitteroecker P. Semilandmarks: A method for quantifying curves and surfaces. Hystrix. 2013;24(1):103–9.

57. Hutton TJ, Buxton BF, Hammond P, Potts HWW. Estimating average growth trajectories in shape-space using kernel smoothing. IEEE Trans Med Imaging. 2003 Jun;22(6):747–53.

58. Evans DM. Elucidating the genetics of craniofacial shape [Internet]. Vol. 50, Nature Genetics. Nature Publishing Group; 2018. p. 319–21. Available from: https://doi.org/10.1038/s41588-018-0068-1

59. Dense Surface Model [Internet]. Available from: https://www.photomodeler.com/downloads/OnlineHelp/index.html#!densesurfacemodel.htm

60. Hutton TJ, Buxton BF, Hammond P. Dense surface point distribution models of the human face. Proc Work Math Methods Biomed Image Anal. 2001;(October 2001):153–60.

61. Çeliktutan O, Ulukaya S, Sankur B. A comparative study of face landmarking techniques [Internet]. Vol. 2013, Eurasip Journal on Image and Video Processing. Springer International Publishing; 2013. p. 13. Available from: https://jivp-eurasipjournals.springeropen.com/articles/10.1186/1687-5281-2013-13

62. Loh P. BOLT-LMM v2.3.4 User Manual. 2019;1–23.

63. Koehrsen W. Introduction to Bayesian Linear Regression [Internet]. 2018. Available from: https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7

64. Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat Commun [Internet]. 2019; Available from: https://doi.org/10.1038/s41467-019-12653-0

PART II
Thesis project

1 Introduction

The face is one of our most important human features, it houses four of our five senses and provides the means to communicate with others. It enables us to show emotions, express ourselves and holds clues regarding one's identity: our age, sex and ancestry. This results in the face being a complex, multipartite structure. Even more so since each face is unique and distinguishable, yet every face is composed of the exact same features: two eyes, a nose, a mouth, cheekbones, a jawline etcetera.

The development of the face starts during the fourth gestational week and follows a precisely-timed path of gene expressions and molecular interactions (1). Postnatally, the face develops following a pattern of periods with steady growth mixed with periods of rapid growth, peaking during puberty (1). The development of the face is influenced by genetic, environmental, epigenetic factors and their interactions (2–4). The influence of the genetic component can be seen in resemblances amongst relatives and the likeness between monozygotic twins. The influence of environment is present from the moment of conception and continues to influence facial morphology during the whole lifetime (1,2,5). Research has found that some regions of the face are more influenced by environmental factors while others are predominantly shaped by genes (1). The genetic influence on a phenotype, in this case facial morphology, is quantified as the heritability; the proportion of variation in a phenotype explained by genetic variance (2,6). Djordjevic et al. investigated the genetic contributions to facial morphology, and found that genetic component explain up to 70% of phenotype variance, with some regions being more susceptible to genetic influences than others (2).

While it is known that a large portion of phenotypic variance is explained by genetics, the underlying genetic architecture is mostly unknown (1,7). For a long time, the focus of genetic research was on the genetics of diseases and not so much on the genetics of normal-range variation (1,2). Recently, this field has gotten more attention and a number genetic loci underlying facial features have been identified through multiple studies (8–19). Understanding the genetic architecture is relevant for several reasons. First, understanding the genetics of normal-range variation may aide in the comprehension of the genetic basis of diseases (20). Second, it could help unravelling the contributions of genetic and environmental factors on the face.

The high heritability and genes associated with facial morphology are part of identifying the genetic architecture of the facial features, however these don't give insight in how the facial features are shaped by genetics. To investigate the relationship between the genotype and the phenotype, predictive modelling can be used. Predictive modelling is a term used to describe genetic prediction models: Tools that predict the phenotype of an individual based on the DNA. For some traits, successful predictive models are available, such as the colour of the eyes, skin and hair (21).

In this thesis project, the possibility of predictive modelling of facial features from DNA was explored. The prediction of facial features is a complicated task due to three main reasons. First, the facial morphology is a complex trait since it is a collection of different features and there is not a uniform approach to capture variation is the phenotype. Second, for predictive modelling a large dataset is needed for a model to find associations between variations in genotype and phenotype. Third, the heritability of the face is known, however the exact mechanisms are not; the contribution of individual genes to the phenotype and gene interactions are unknown.

The aim of the thesis project was to explore the possibility of predictive modelling of facial features and to investigate factors that influence the prediction of facial features. Specifically, the influence of age and gender on the prediction were investigated. As the body ages, structural changes in tissues

occur, leading to morphological changes in the face (22). In addition, exposure to environmental factors such as UV-light, nutrition and toxins (such as alcohol and cigarette smoke) contributes to aging of the face (22). Thus the contribution of the environmental component on facial features increases due to long-term exposure to environmental factors. To investigate this, the predictive modelling was applied to two separate datasets: one with a study population of children and one with a study population of adults.

2 Methods & Materials

2.1 Study design

In this chapter I will discuss the methods used for my thesis project. First, I will describe the outline of the methods and discuss the different steps of the project. In the subsections I will dive deeper into each step and give a detailed description. The project pipeline can be divided into five main steps: data pre-processing, dimensionality reduction, the genetic prediction model, 3D facial shape reconstruction and the similarity measure. The complete pipeline of this project is visualized in Figure 1, where each step is enclosed by a dashed rectangle with the number of the step in the upper left corner.

The first step was pre-processing of the datasets. For this study, data was used from two independent studies: Generation R and the Rotterdam Study. From both studies two sets of data were utilized: genetic data and phenotype data. The genetic data consisted of single nucleotide polymorphisms (SNPs) for each dataset. The phenotype data were 3D facial meshes composed of 5023 vertices. The high-dimensionality of the phenotype data is not desirable for a prediction model (23). Liu et al. were able to decrease the dimensionality of the phenotype data from 5023 vertex per subject to 200 points using an auto-encoder(5). These 200 points are referred to as endophenotypes. The endophenotypes computed by Liu et al. were used as the phenotype data for the prediction model (5).

The third step in the pipeline is the genetic prediction model. For this project the LDAK-Bolt-Predict model was used. The LDAK-Bolt-Predict model (LDAK model) is a linear regression model where the heritability contributed by each SNP is computed beforehand based on a specified heritability model (24). In this way it differs from other existing models where the effect size of each SNP is assumed to be constant. The prediction model was trained on a subset of the samples and the remainder was used for testing. For testing, the endophenotypes of the individuals were predicted based on their genotypes using the trained LDAK model, i.e. the models outputted two hundred endophenotypes for each subject in the test set.

Step four encompasses the reconstruction of the face from the 200 endophenotypes to a 3D facial mesh of 5023 vertex. For the reconstruction Liu's framework was used (5). After step four, there is a predicted 3D facial mesh for each subject from the test set.

Fifth, to evaluate the performance of the prediction models, a similarity measure was computed between the predicted 3D facial mesh and the original 3D facial. As mentioned in the introduction, the face is a very complex structure, because of the presence of the eyes, nose and mouth and because of the different curves and surfaces such as the cheekbones, forehead and jaw. To evaluate the performance, different similarity measures were computed as well as different methods for performance evaluation.

In the following subsections, I will go more into the details of each step of the pipeline.

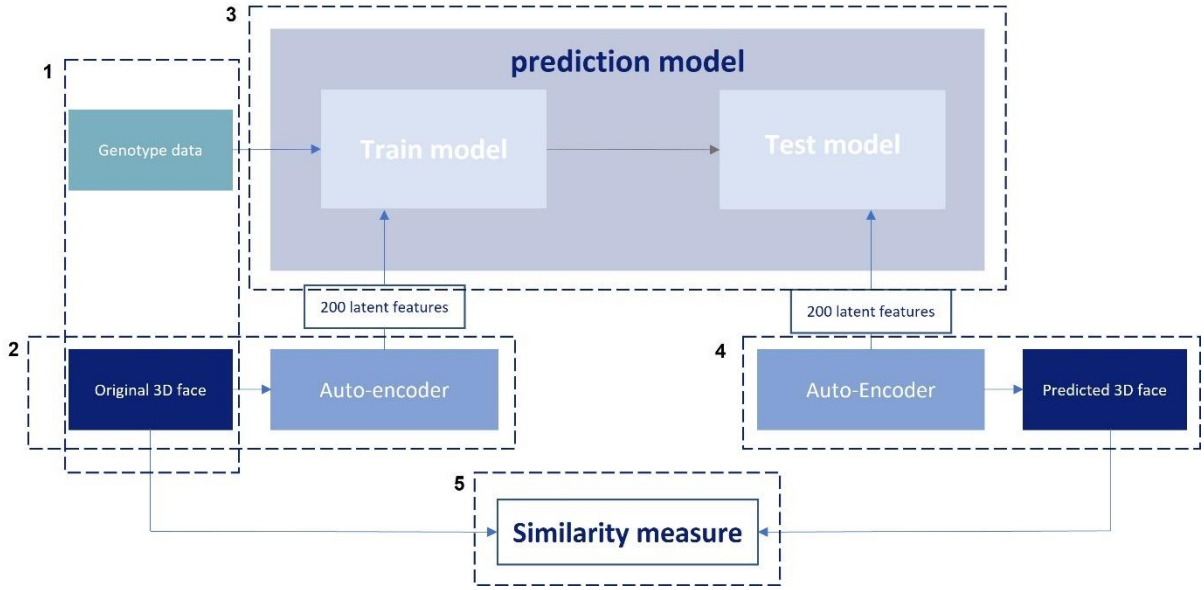

Figure 1: This contains the visualized pipeline of this project. The project can be categorized into five steps, each step is marked by a dashed rectangle with the number of the step in the upper left corner.

2.2 Study population and data pre-processing

During this project, genotype and phenotype data were used from two different studies. Part of the research objective is to evaluate the influence of age on the prediction of facial features from DNA. The two studies have very different populations, the study population of Generation R are children who are followed from foetal life until young adulthood (25), and that of the Rotterdam Study are adults over the age of 45 (26).

2.2.1 Generation R

Generation R is a multi-ethnic prospective cohort study that investigates the growth, development and health of children in Rotterdam. In total, 9778 women with a delivery data between April 2002 and January 2006 were included in the cohort (27). The Generation R study aims to identify genetic and environmental causes in normal and abnormal development, growth and health of a child (27). From foetal life until young adulthood data is collected from the parents and children, in the form of, amongst other things, questionnaires, physical examinations, behavioural observations, biological sampling (27). The main objective of the study is to uncover why some children have an optimal development whilst others don't or show suboptimal development (4).
For this thesis project, genetic data and phenotype data gathered in the Generation R study were used. As mentioned above, the phenotype data used for this project are the endophenotypes derived from 3D facial images.

2.2.2 The Rotterdam Study

The Rotterdam Study is a prospective cohort study as well, the study started in 1990 in the Ommoord district in Rotterdam (29). The study was initiated in response to the demographic changes leading to aging populations worldwide, as it was foreseen that this would lead to increasing numbers of chronic illnesses in mid- and late-life (29). The objective of the study is to uncover causes of diseases, identify risk factors and find targets for preventive interventions (29). In the first and second cohort, the study included participants over the age of 55 years, the third cohort participants aged 45 years and above

were included as well. The fourth cohort, started in 2016, includes participants from the age of 40 years (29). A thorough examination is performed upon entry of the study, the baseline, and every three to six years re-examination takes place, data is collected in the form of interviews, imaging, and biological sampling (29). For this thesis project, genetic data and phenotype data gathered in the Rotterdam study were included.

2.2.3 Data formats

Genotype

The genetic data used for the prediction were single nucleotide polymorphisms (SNPs). The SNPs were obtained using genotyping arrays, however the studies used different arrays to detect the SNPs. The SNPs from Generation R were obtained using genome wide Illumina 610 or 660 Quad chips (30). For the Rotterdam study different arrays were used between the cohorts. For the first and second cohort (RS-I, II) a 550K Illumina array was used and for the third cohort (RS-III) a 610K Illumina array was used (31). For both studies, the non-imputed SNPs were used.

Phenotype

The phenotype data used for this project are the endophenotypes derived from 3D facial meshes. The 3D facial meshes were computed from 3D facial images. The 3D facial images of both study populations were captured using a 3dMD camera system. Participants were asked to maintain a neutral expression for the 3D images. The 3D images are favourable to 2D image since they preserve depth information. Moreover, the influence of head pose and lighting is minimized. The 3D facial meshes are all composed of the same number of vertices and triangles and aligned to all be in the same position and orientation. Vertices are points in 3D space, between the vertices triangles are formed which results in a 3D surface mesh. Figure 2 visualizes how the 3D facial mesh is computed from the 5023 vertices. Because all meshes are aligned, the vertices have the same positions in different individuals (e.g. vertex *i* is the tip of the nose for all individuals). This makes that vertex between subjects are pairs and one-to-one comparison is possible.

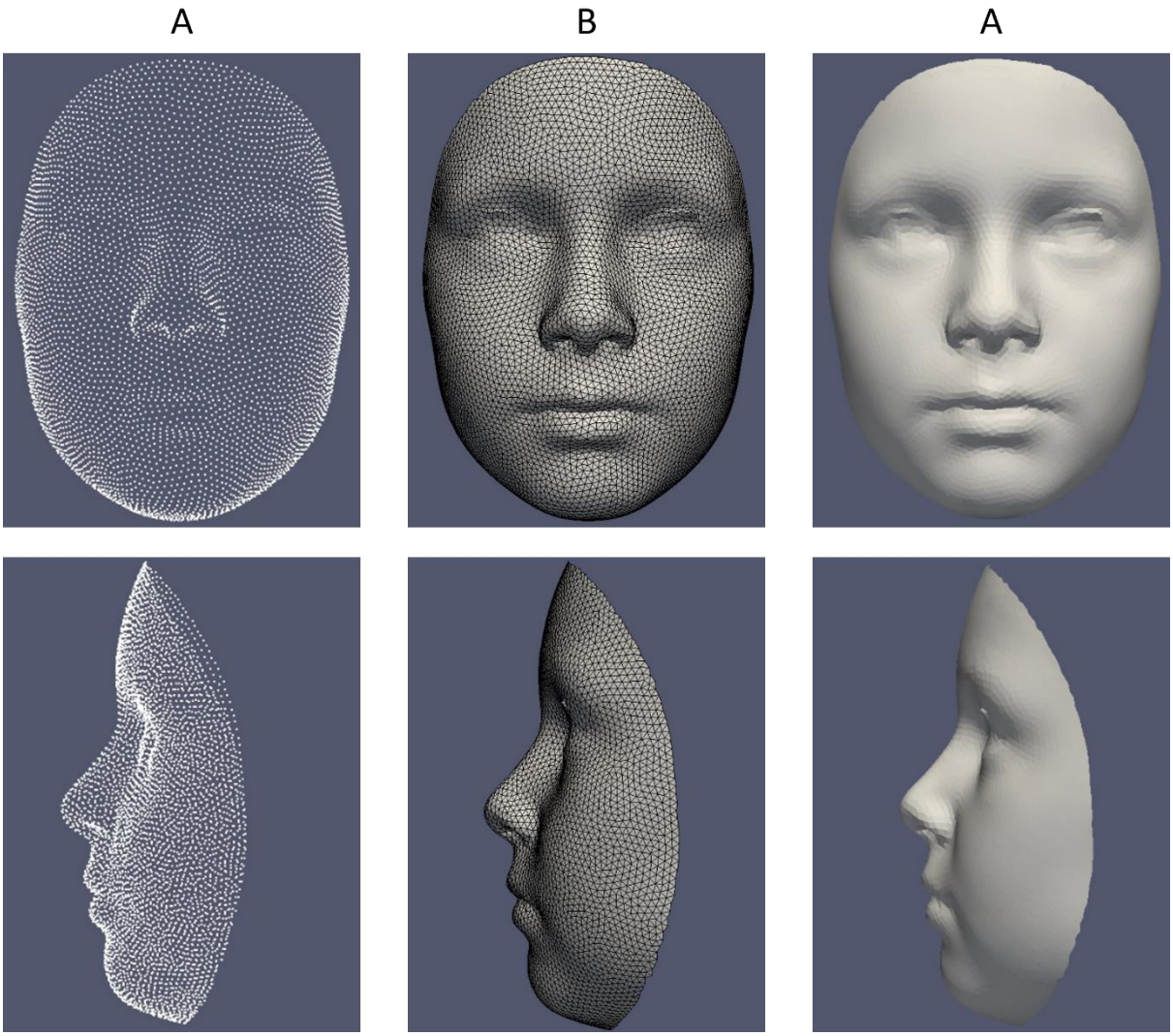

Figure 2. The 3D facial images are represented by 5023 3D points (vertices). The 3D facial images are visualized as a 3D point cloud in column A. To transform the separate vertices into a connected mesh. Triangles are computed between the vertices. This results into one consecutive mesh, like column B shows. When removing the edges of the triangles a smooth mesh remains such as in column C.

2.2.4 Data inclusion

From the Generation R and Rotterdam study participants were included in the dataset for this project if both genetic and phenotype data was available. In case of siblings in the dataset, only one sibling was included in the dataset. Since the inclusion of siblings can lead to correlations based on ancestry and not due to causality, which influences the performance of the genetic prediction model (32). Because the Rotterdam Study consists of different cohorts, the genetic data of each cohort were merged into one dataset.

For Generation R data from subjects at the age of 9 was included, the first time they got their 3D facial images taken. It was assumed the influence of environment is smaller at a younger age. Therefore the 3D facial images were used that were obtained at the youngest age. For the Rotterdam Study all available subjects were included, there was no upper or lower age threshold.

2.3 Dimensionality Reduction

As mentioned earlier, the complexity of facial morphology poses a challenge for research regarding facial morphology. During this study, an autoencoder was utilized to manage this obstacle. An auto-encoder is a type of neural network that aims to copy the input of the model to the output. At first glance, this does not seem particularly useful. However, what makes an auto-encoder special is that they can be restricted internally, which forces the network to learn the useful properties of the data (33). This restriction is possible because of the bottleneck in the architecture of the network (34), as visualized in Figure 3. An auto-encoder consists of two parts: an encoder and a decoder (33). The encoder is able to map the input into the lower dimensional code, this lower dimension is called the latent space (35). Thus, the latent space is a compressed representation of the input. The decoder reconstructs the input of back from the latent space to the dimensionality of the input data, thus the output is a reconstruction of the input. The auto-encoder learns through minimizing the reconstruction error, also referred to as the loss function, this is a measure of error between the input and the output.

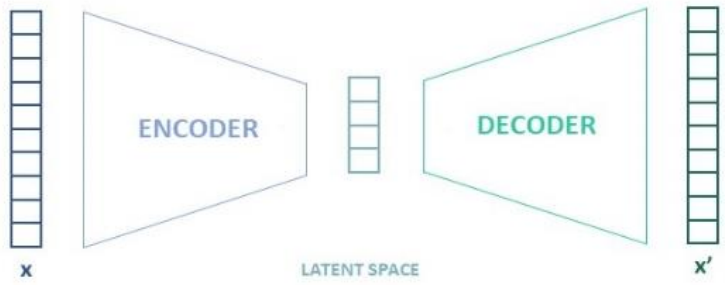

*Figure 3: The general architecture of an auto-encoder. The input **X** is fed into the encoder, which compresses the input into a lower dimensionality, the latent space. The decoder resamples the latent space into the output **X'**, a reconstruction of the input.*

The lower dimensionality of the latent space can be beneficial for the performance of a tasks like classification and a smaller dimensionality requires less computational power and runtime (33). Moreover, because of the compression of the input data the latent space contains the important information or features that represent the input data.

The endophenotypes used in this study were computed using Liu's framework, a 3D graph auto-encoder (5). Figure 4 shows the architecture of the Auto-encoder used by Liu. The latent space of the framework was set to 200 latent features, these latent features are the endophenotypes. The authors investigated the influence of different latent space sizes on the reconstruction error and a trade-off was made between the reconstruction error and the dimensional complexity, setting the latent space size to 200 features (5).

Liu's framework is capable to compress the 3D facial shapes, consisting of 5023 vertices, into 200 latent features. The decoder of the framework can reconstruct the 3D facial mesh of 5023 points from the 200 latent features. The auto-encoder is specific for a study population, thus it was trained separately for the Generation R dataset and the Rotterdam Study data set. For the Generation R dataset the auto-encoder was trained on ~9000 subjects. For the Rotterdam Study the network was trained on ~5700 subjects.

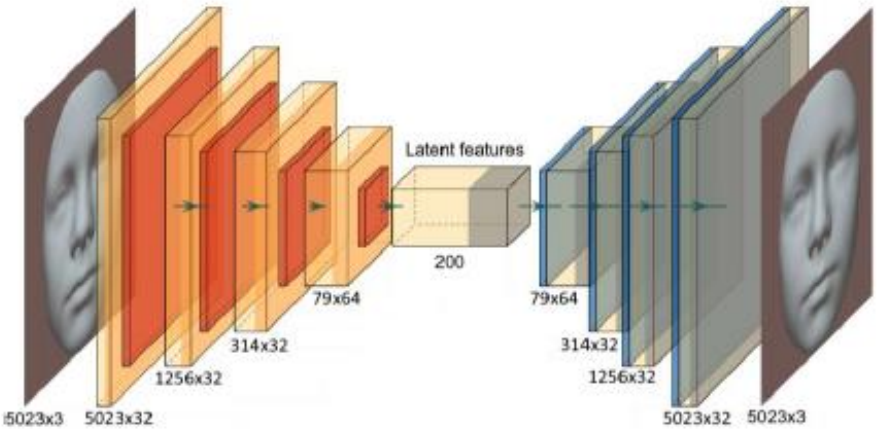

Figure 4. The architecture of the auto-encoder used by Liu et al. The orange layers are the encoder, the blue layers are the decoder. The dimensionality of the input image is gradually decreased from 5023x3 to 200 latent features (5). The decoder resamples the latent features to the dimensions of the input image.

2.4 Genetic prediction model

To perform the genetic prediction of facial features, the LDAK-Bolt-Predict model (LDAK model)was used during this project (24). This model is part of the LDAK-software, which contains tools for genetic prediction of complex traits, either from individual-level data or summary statistics. The LDAK software distinguishes itself from other available tools by allowing the user the specify the heritability model, which determines the expected heritability of each SNP (24). In contrast to most prediction tools, where a constant heritability is used for all SNPs. The SNP heritability is the proportion of phenotypic variance that is explained by SNPs (6). The LDAK software was tested for several complex traits (binary, continuous and ordinal) and all eight tools outperformed existing prediction models which use the GCTA model (24). The LDAK model is a multilinear regression models described by Equation 1, Where X_j is the genotype for SNP *j* and β_j is the effect size for SNP *j* (24).

$$[1] \qquad E[Y] = X_1 \cdot \beta_1 + X_2 \cdot \beta_2 + \ldots + X_m \cdot \beta_m = X \cdot \beta$$

Equation 1: the linear model of the LDAK prediction model

Since two different datasets were used, two separate LDAK-models were trained and tested: one for Generation R and one for the Rotterdam Study. Before running the model, the expected heritability contributed by each SNP, $E[h_j^2]$, was computed. This was done with the heritability model, for this project the BLD-LDAK heritability model was used (36). This model determines the $E[h_j^2]$ based on 66-parameters: 64 functional annotations, local levels of linkage disquilibrium and the minor allele frequency (MAF) (24). The computation of the expected heritability of each SNP was computed using the LDAK software (36). For the prediction model, the LDAK-Bolt-Predict model was used, which is recommended for individual-level genotype and phenotype data. Each of the tools from the LDAK software uses a different prior distribution for the SNP effect size β_j. For the LDAK-Bolt-predict model the effect size β_j is based on the $E[h_j^2]$ and two pre-defined variables. During training the effect size β_j is updated iteratively.

Age and sex were added as covariates to the model. The model was trained on ninety percent of the datasets, the remaining ten percent was used as the test set. The LDAK model performs multilinear regression for each endophenotype individually, thus no interaction between the endophenotypes is possible and they are seen as independent phenotypes. For each endophenotype the R^2 was computed.

2.5 Similarity Measure

During this project, several approaches were used to evaluate the performance and to compute the similarity measure. In section 2.5.2 the different methods for the evaluation of the performance and similarity metrics are described. However, before analysing the performance of the genetic prediction method, the reconstruction error of the auto-encoder was computed. Both the ground truth as the predicted face are represented by a mesh made up of 5023 vertices and 9851 triangular faces (28). The 3D facial image was captured with the 3D camera system and was translated into the 3D mesh. However, the predicted faces were reconstructed into a 3D mesh from the 200 endophenotypes. This was done with the decoder from the auto-encoder, as described in section 2.3. The aim of the auto-encoder is to minimize the reconstruction error between the input and the output, however some loss of information is inevitable due to the bottleneck architecture. The reconstruction error was computed to investigate the extent to which differences between the ground truth and predicted face are due to the reconstruction error and not as a result of the prediction accuracy.

2.5.1 Reconstruction Error

Since the reconstruction error influences the similarity measure between the prediction and the ground truth, it is necessary to know the reconstruction error of the auto-encoder. The reconstruction error was evaluated using the Generation R test set. The ground truths were fed into the encoder and reconstructed with the decoder without making changes the endophenotypes in the latent space. To quantify the reconstruction error the mean Euclidean distance was calculated for every vertex-pair using Equation 2, where $v^i$is the vertex in the ground truth and v_r^i is that same vertex in the reconstructed face. Thereafter, the total mean error was computed, i.e. the mean of the 5023 vertex errors.

[2] $$d(v^i, v_r^i) = \sqrt{(x - x_r)^2 + (y - y_r)^2 + (z - z_r)^2} \; for\; i = [1, \dots, 5023]$$

Equation 2

The reconstruction error for each vertex was visualized in a template face with a heatmap.

2.5.2 Similarity Metrics

For the similarity measure a global and a learning-based metric were computed. The metrics differ in the way the total error between two 3D facial meshes in computed. The global similarity metric consisted of calculating the mean squared error (MSE) between two 3D facial meshes. For this metric, Equation 3 was used. Where v_{ref}^i is the vertex in the reference face and v^i is the corresponding vertex in the test set. The error between two vertices, $v_{ref}^i - v^i$, is the Euclidean distance (see Equation 2).

[3] $$MSE = \frac{1}{n} \sum_{i=1}^{n} (v_{ref}^i - v^i)^2, for\; i = [1, \dots, 5023].$$
$$Where\; n\; is\; the\; total\; number\; of\; vertices$$

Equation 3

For the learning-based metric, a weight between zero and one was assigned to each vertices-pair. This was done with the idea that some vertices are more predictable than others and thus are more important for the prediction evaluation. To compute the weights, the test sets were split into two subsets: 70% of the samples were used for training the weights and 30% for testing. Equation 4 is the formula of the computation of the MSE using the learning-based metric. The error between the two vertices, $v_{ref}^{i} - v^{i}$, is the Euclidean distance between them (see Equation 2)

[4] $$MSE = \frac{1}{n}\sum_{i=1}^{n}(w_i * (v_{ref}^{i} - v^{i}))^2 , for\ i = [1, \dots, 5023].$$
$Where\ n\ is\ the\ total\ number\ of\ vertices\ and\ w_i\ is\ the\ weight\ assigned\ to\ vetrices - pair\ i.$
Equation 4

The weights for each vertex-pair, w_i, were trained using a PyTorch nn.Linear module. This is a single layer network with 5023 inputs, i_i, one for each vertex, one neuron and a single output. The inputs are the vertex-pair errors. The output is the total error between the prediction and the ground truth, in the ideal scenario this would be zero (i.e. the prediction is perfect). The architecture of the linear module is visualised in Figure 5.

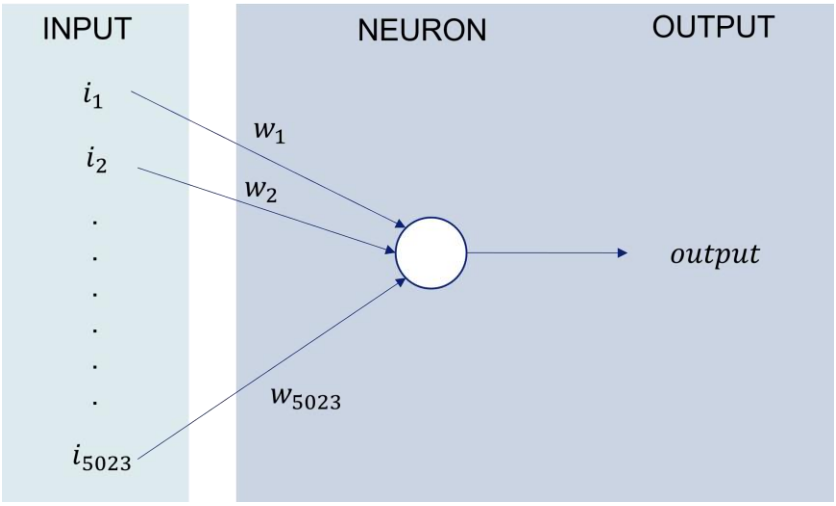

Figure 5: network architecture for the learning-based metric. The input layer contains an input for each vertex-pair, a weight is assigned to each input.

Initially, all inputs had the same weight, this was set to one. During training of the model the loss between the output and the desired output is calculated and used to compute the gradient. The weights are updated using stochastic gradient descent with a learning rate of 0.1 and the number of epochs was 5. The weights trained in the model are applied in Equation 4.

2.5.3 Performance evaluation metrics

In addition to the two different similarity metrics, two different metrics for performance evaluation were used. First, to investigate whether the predicted face was most similar to the corresponding ground truth or another individual in the test set, a similarity measure was computed between each prediction and all ground truths in the test set. For the similarity measure, both the global and the learning-based metric were computed as described by Equation 3 and Equation 4. Thus for each predicted face there was a similarity measure to all ground truths. These were sorted in an ascending order, the ground truth with the smallest MSE was deemed as the 'closest' to the prediction and so on. Next, the rank of the corresponding ground truth was determined (e.g. if the MSE between the

prediction and ground truth of subject *i* was the second smallest, it would be ranked second). The ranking method is visualized in Figure 6.

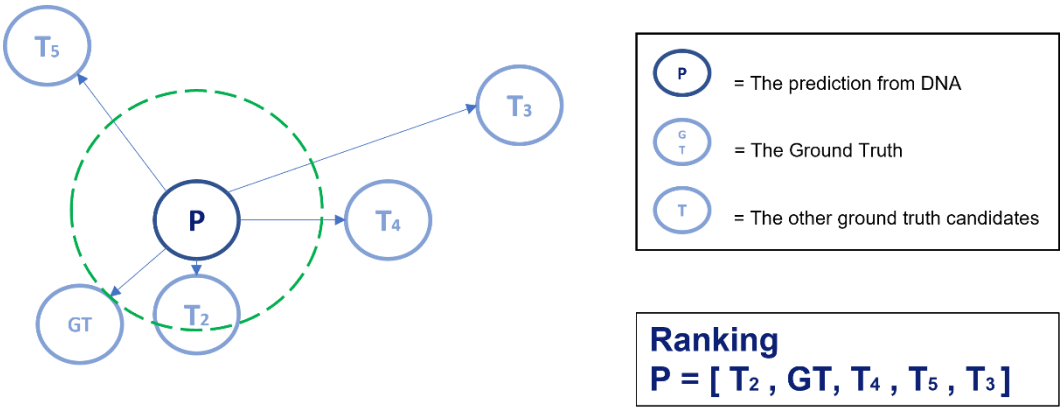

Figure 6. Visual representation of the second similarity measure.

Thus, the first performance evaluation focusses on identification of the correct individual based on the predicted face from the genetic prediction model.

Second, there is the possibility that individuals in the dataset are in close resemblances to each other. In this case, there is the probability that the predicted faces are in close resemblance to other subjects in the test set. To get an insight in this scenario, a second performance measure was computed. The similarity measure between all ground truth faces in the test set was computed, using both similarity metrics. The similarity metrics were ranked the same manner as described above, in ascending order. In addition, the similarity measure between each ground truth and their prediction was computed using both similarity metrics as well. Next, the position in the ranking of the similarity metric between the ground truth and prediction was determined. Figure 7 visualizes the ranking process for this similarity measure.

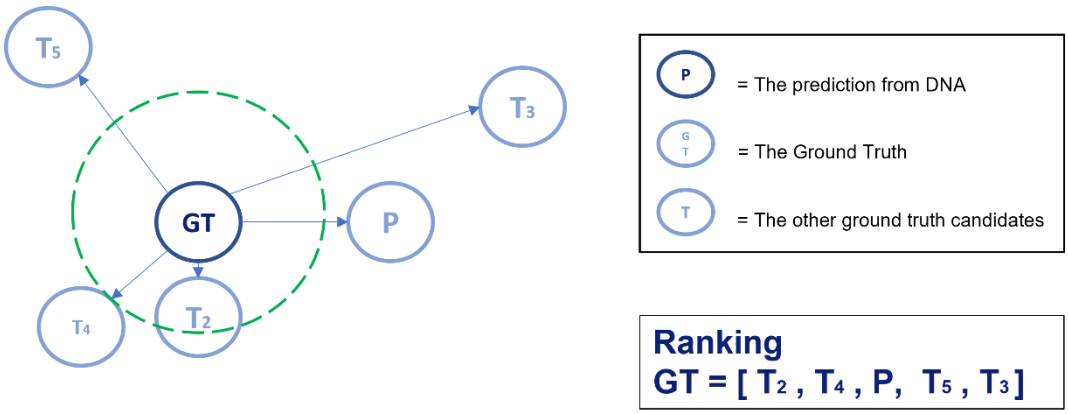

Figure 7. Visual representation of the method for ranking used in similarity measure three.

The performance evaluation was performed using the complete test sets. In addition, the test sets were divided into sub-sets based on sex and ethnicity. The latter could solely be done for the Generation R dataset, from the Rotterdam Study information on ethnicity was not available.

2.5.4 Accuracy Ratio

For both prediction evaluation metrics, the accuracy ratio (AR) was computed. The AR allows a quantitative comparison between the different datasets. It is the ratio of the improvement of the prediction model over a random model (the baseline) to the performance improvement of a perfect model . The AR is a value between zero and one, where zero indicates no improvement over the baseline and one a perfect model. The concept of the AR is visualized in Figure 8. The AR is calculated using Equation 5.

[5] $$Accuracy\ Ratio = \frac{B}{(A+B)}$$

Equation 5: Formula to compute the Accuracy Ratio

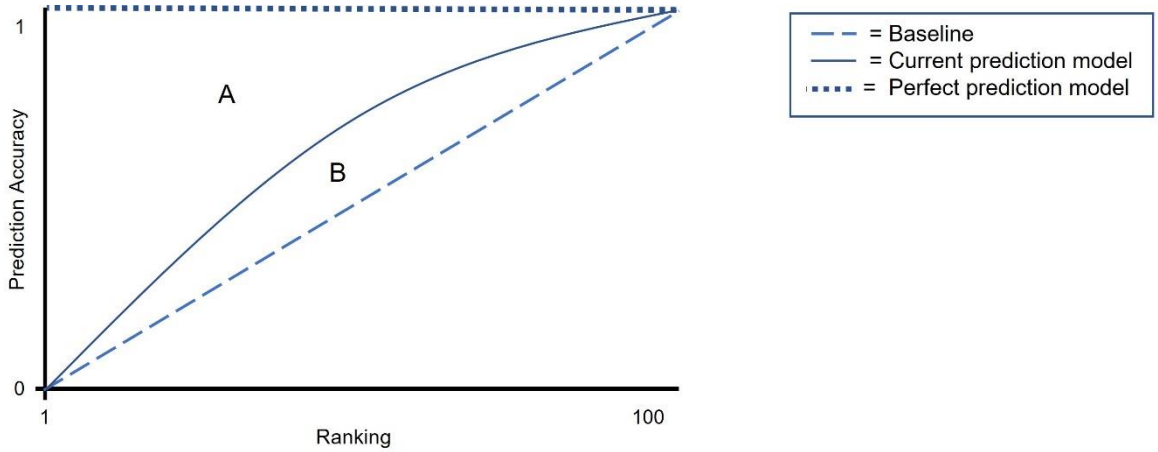

Figure 8: Visualisation of the accuracy ratio

The baseline is the chance the correct individual is in the top-n when randomly picking n-individuals from the dataset. To compute the baseline, Equation 6 was used.

[6] $$Baseline = \frac{n}{m}\ where\ n\ is\ the\ ranking\ and\ m\ is\ the\ total\ number\ of\ subjects$$

Equation 6: Formula used to determine the baseline

4 Results

4.1 Data pre-processing

Before training of the prediction model, data was included from the Rotterdam Study and Generation R. From the Rotterdam Study 3197 subjects were included after removal of relatives. The genotype data of consisted of 545.018 SNPs, obtained with Illumina genotyping arrays (31). Figure 10a shows the flowchart for the data selection. From Generation R 2774 subjects were included, after removal of relatives. Figure 10b shows the flow diagram of the subject exclusion for Generation R. 518.243 SNPs were used for the genetic prediction for Generation R. Both datasets were split into a training and test set, 90% of the data was used for training of the LDAK prediction model and 10% was used for testing the model. The split of the data into the training and test set was randomized. The characteristics of the training and test sets can be found in Table 2.

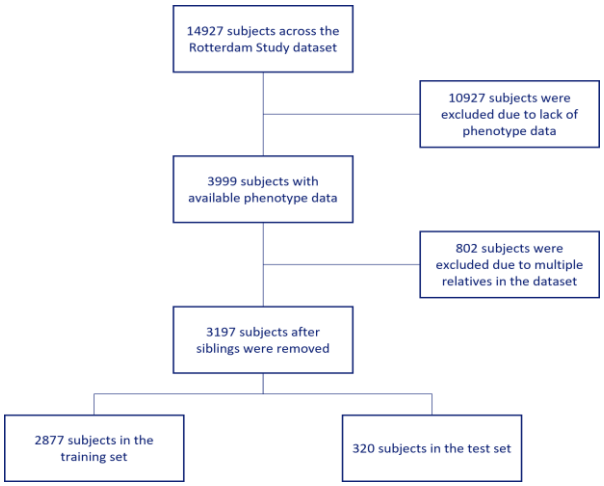

Figure 10a: Flowchart of study data selection for the Rotterdam Study. Flow diagram shows the exclusion and inclusion of subjects from the Rotterdam Study.

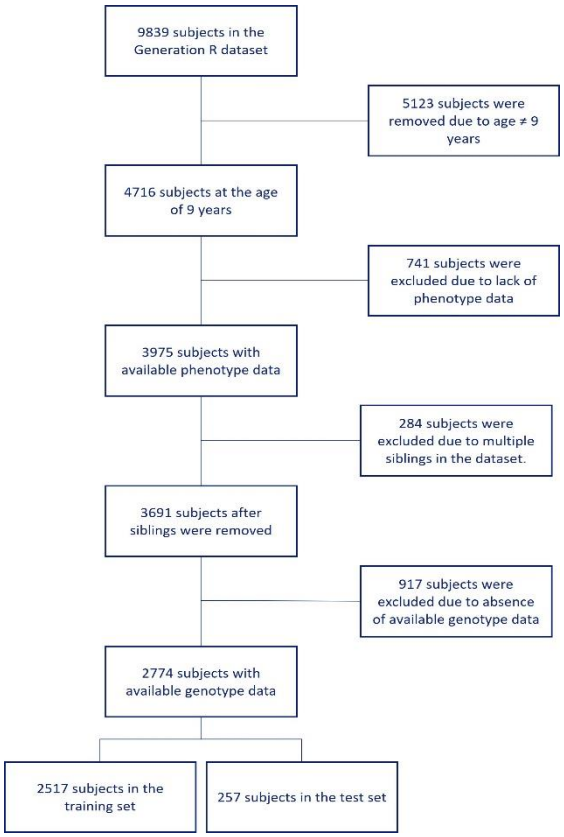

Figure 10b: Flowchart of study data selection for Generation R. The diagram shows the exclusion and inclusion of subjects from the Generation R study.

		Generation R	Rotterdam Study			Generation R	Rotterdam Study
Training set	*Total, n*	2517	2877	**Test set**	*Total, n*	257	320
	Sex, n (%)				*Sex, n (%)*		
	Male	1219 (48)	1255 (44)		Male	124 (48)	144 (45)
	Female	1298 (52)	1622 (56)		Female	133 (52)	176 (55)
	Age (year), mean ±SD	9,81 ±0,35	70,23 ±9,56		*Age (year), mean ±SD*	9,81 ± 0,29	70,37 (9,42)
	Ethnicity, n (%)				*Ethnicity, n (%)*		
	Western	1553 (62)	-		Western	150 (58)	-
	Non-western	964 (38)	-		Non-western	107 (42)	-

Table 2: Subject characteristics of the training and test sets for the Generation R and the Rotterdam Study.

4.2 Reconstruction Error

The endophenotypes used for the prediction were reconstructed to a 3D facial mesh using the auto-encoder described in section 2.3. The reconstruction error of the auto-encoder was evaluated by running the test-set (n = 257) through the auto-encoder without altering the endophenotypes in the latent space. The mean error for each vertex-pair is visualized in Figure 11 on a template facial mesh. The error, Euclidean distance, was calculated between each vertex-pair for all 257 individuals (i.e. 5023 pairs per individual, for all 257 individuals in the test-set). Figure 12 is a histogram of the errors for all subjects. Thereafter, the overall mean error was computed. The overall mean error found was 0,43 mm with a standard deviation of 0,08 mm.

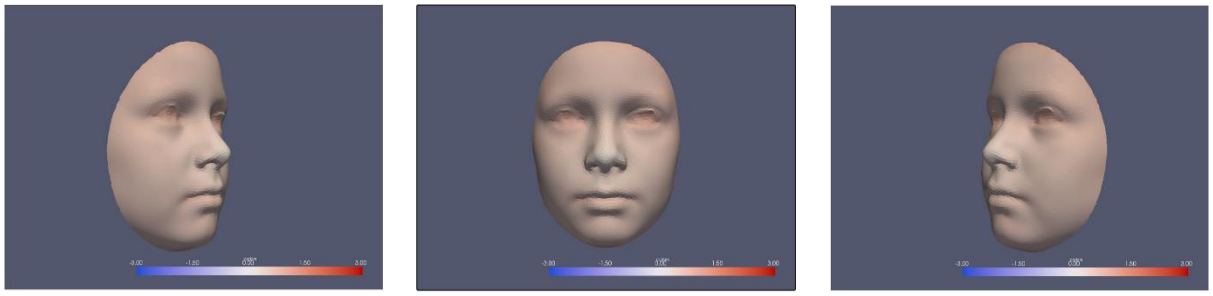

Figure 11: Heatmap of the mean error for each vertex-pair visualized on a template face

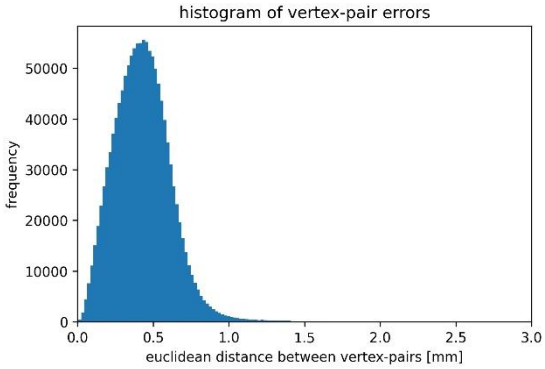

Figure 12: Histogram of all errors between the vertex-pairs of the ground truth and reconstructed facial meshes.

4.3 Performance LDAK-model

To predict the endophenotypes, multilinear regression was performed for each endophenotype individually. Figure 13 displays a scores plot for one of the endophenotypes. For each endophenotype, the R^2 was computed. The R^2 are visualised using a heatmap in Figure 14. The top left is the R^2 for endophenotype one, the second endophenotype is on the right from endophenotype one and so one.

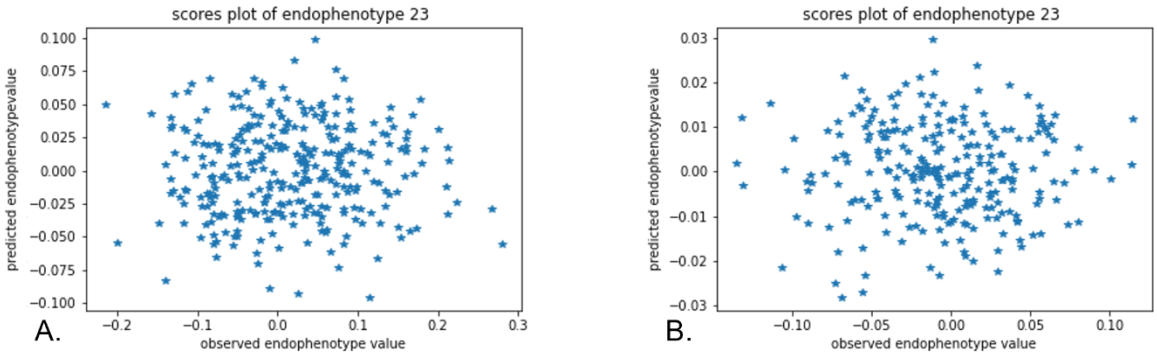

Figure 13: The scores plots for endophenotype 23. A) Scores plot from the Rotterdam Study model. B) The scores plot from the Generation R model

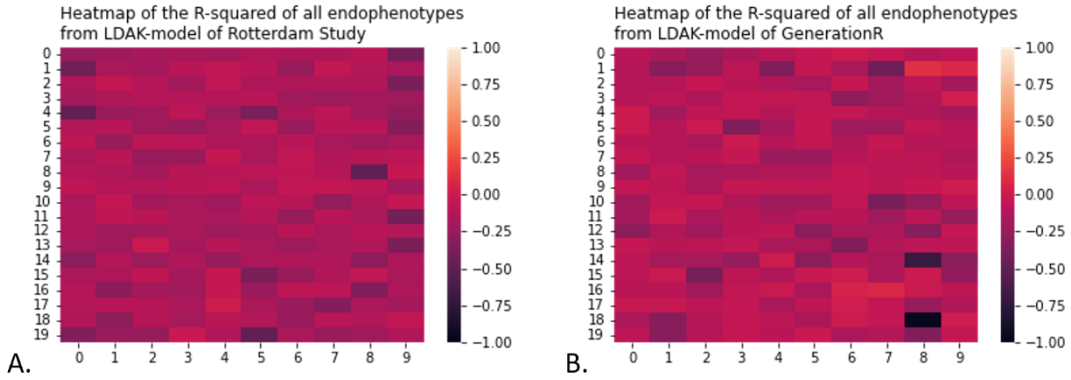

Figure 14: heatmaps of the R-squared for each endophenotype. A) contains the r-squared for the Rotterdam Study endophenotypes. B) contains the heatmap for the Generation R endophenotypes.

4.4 Performance of the prediction

As described in section 2.5 two different similarity metrics were used as well as two metrics for performance evaluation. First, I will discuss the results obtained using the first performance evaluation metric (described by Figure 6). Where the ranking is performed from the predicted face to all ground truths in the test set. I will include the results for both the global similarity metric and the learning-based similarity measure. Second, I will describe the results found using the second performance evaluation metric, for both the global as the learning-based similarity metric.

4.4.1 Prediction evaluation measure one

For Generation R there were 257 subjects in the test set, for the Rotterdam Study there were 320 subjects in the test set. Figure 15, contains the accuracy plots for the prediction with the BLD-LDAK heritability model for the global and learning-based metrics for the Generation R test set. Figure 16 shows the accuracy plots for the Rotterdam Study. In both figures, the blue line indicates the baseline accuracy.

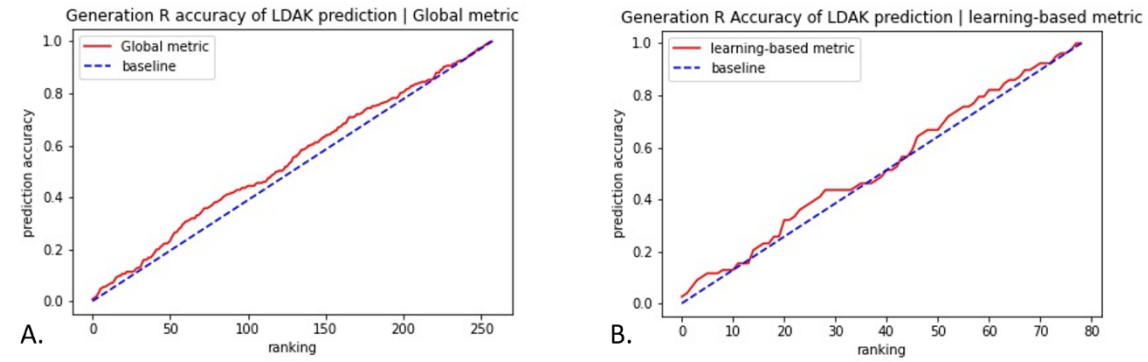

Figure 15: Accuracy plots of the LDAK-prediction model for Generation R. Figure A shows the accuracy plot for the global similarity measure. Figure B displays the accuracy plot of the LDAK-prediction using the learning-based similarity metric

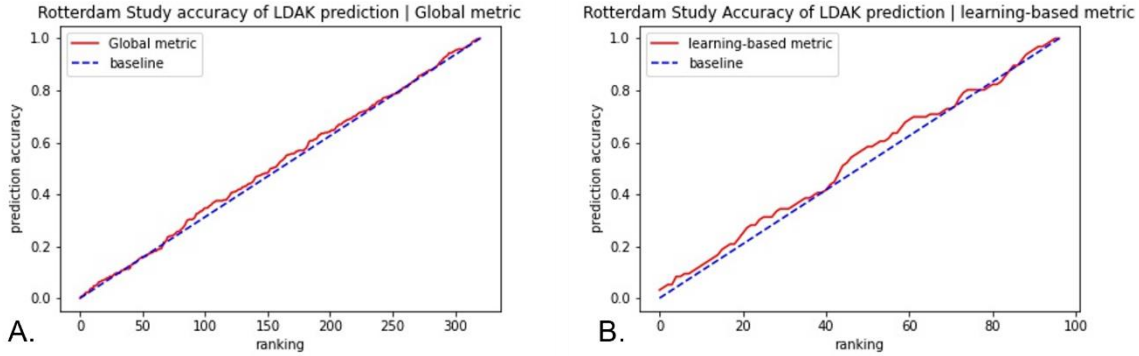

Figure 16: the accuracy plots for the LDAK prediction of the Rotterdam Study dataset. Figure 12A shows the accuracy plot when using the global metric. Figure 12B contains the accuracy plot when using the learning-based metric as the similarity measure.

The test sets used for the accuracy plots in Figure 15 and Figure 16 contained all subjects, there was no division made based on sex or ethnicity. In the next step, the test sets were divided into subsets based on these characteristics. For the Generation R the test set was once separated based on gender and the similarity metrics and performance evaluation were computed. In addition, the test set was divided based on ethnicity, western or non-western, and evaluated. For the Rotterdam Study there is no data available on ethnicity, thus only a division by sex was made. The results of this different sub-groups are visible in Figure 17.

In addition to the accuracy plots, the accuracy ratio was computed. The AR was computed to make comparison between the different datasets, Generation R and Rotterdam Study, and the different sub-groups, sex and ethnicity, possible. The AR's for the different datasets and sub-sets can be found in Table 3

BLD-LDAK heritability model

	Global Metric		**Learning-based metric**	
	Generation R	*Rotterdam Study*	*Generation R*	*Rotterdam Study*
All	0,06	0,02	0,04	0,04
Sex				
Male	0,08	0,02	0,06	0,5
Female	0,06	0,02	0,04	0,06
Ethnicity				
Western	0,04	-	0,02	-

Table 3: Accuracy Ratio for the different datasets and sub-sets.

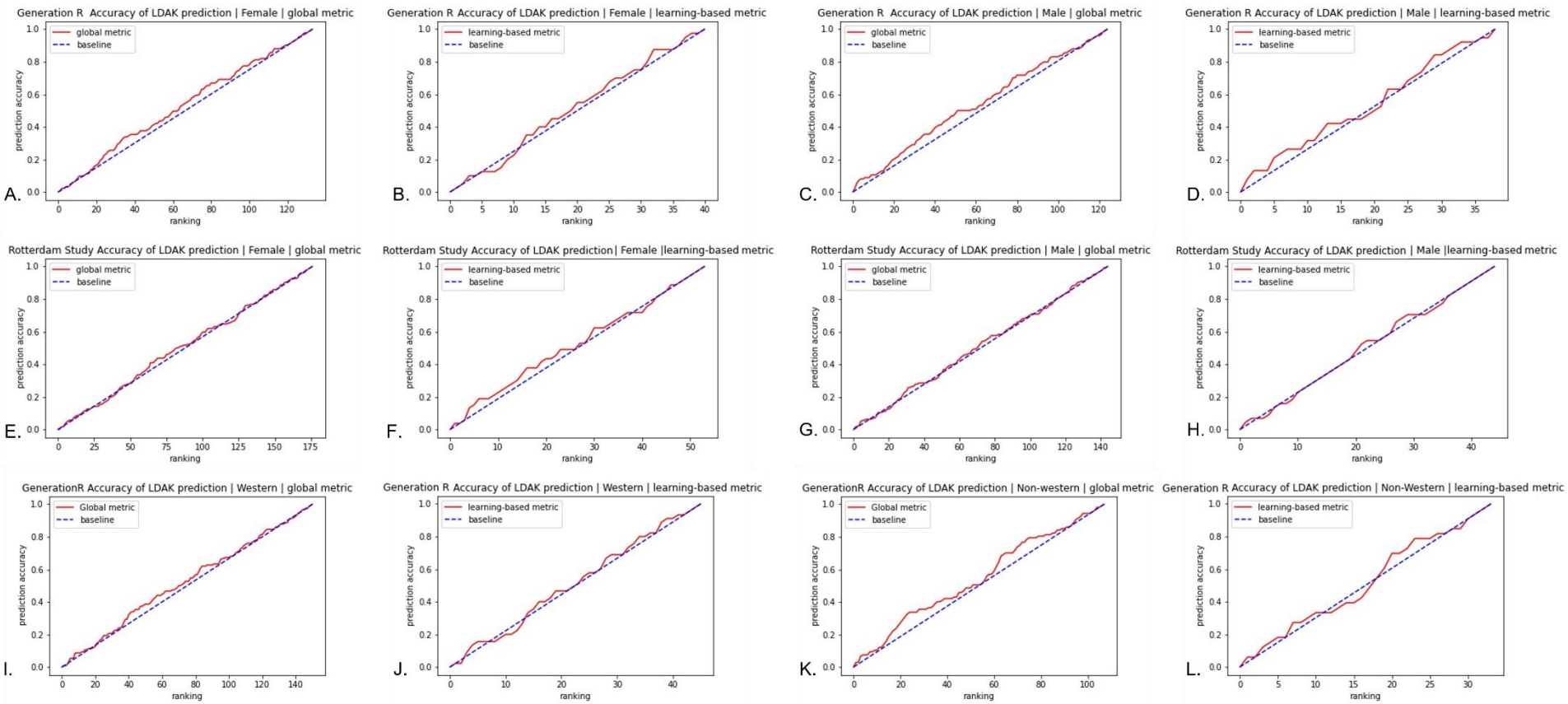

Figure 17: Accuracy plots for the different sub-sets of the Generation R and Rotterdam Study test sets. For each sub-set, both the global metric as well as the learning-based metric were computed. The top row (A-D) are the accuracy plots for the Generation R test set split by sex. The middle row (E – H) are the accuracy plots for the sub-sets of the Rotterdam Study. The bottom row (I – L) contains the accuracy plots for the Generation R data split by Ethnicity.

4.4.2 Prediction evaluation measure two

For the second prediction evaluation measure the similarity measure between the prediction and the ground truth and the similarity measure between all ground truths in the test set was calculated. These were sorted in ascending order and the position of the predicted face was determined based on the value of the similarity measure, as visualized in Figure 7. As in section 4.4.1, the similarity was computed using the global metric as well as the learning-based metric.

For the visualisation of these results, a heatmap was computed of the MSE between the ground truths. A high MSE means a low similarity between subjects. The rank of the predicted face was marked with a white square. There was chosen to use a heatmap to visualize the magnitude of the similarity measures between the ground truth and to get insight whether they were very diverse or close to each other. Figure 18 shows the heatmaps for the complete test sets of Generation R and the Rotterdam Study.

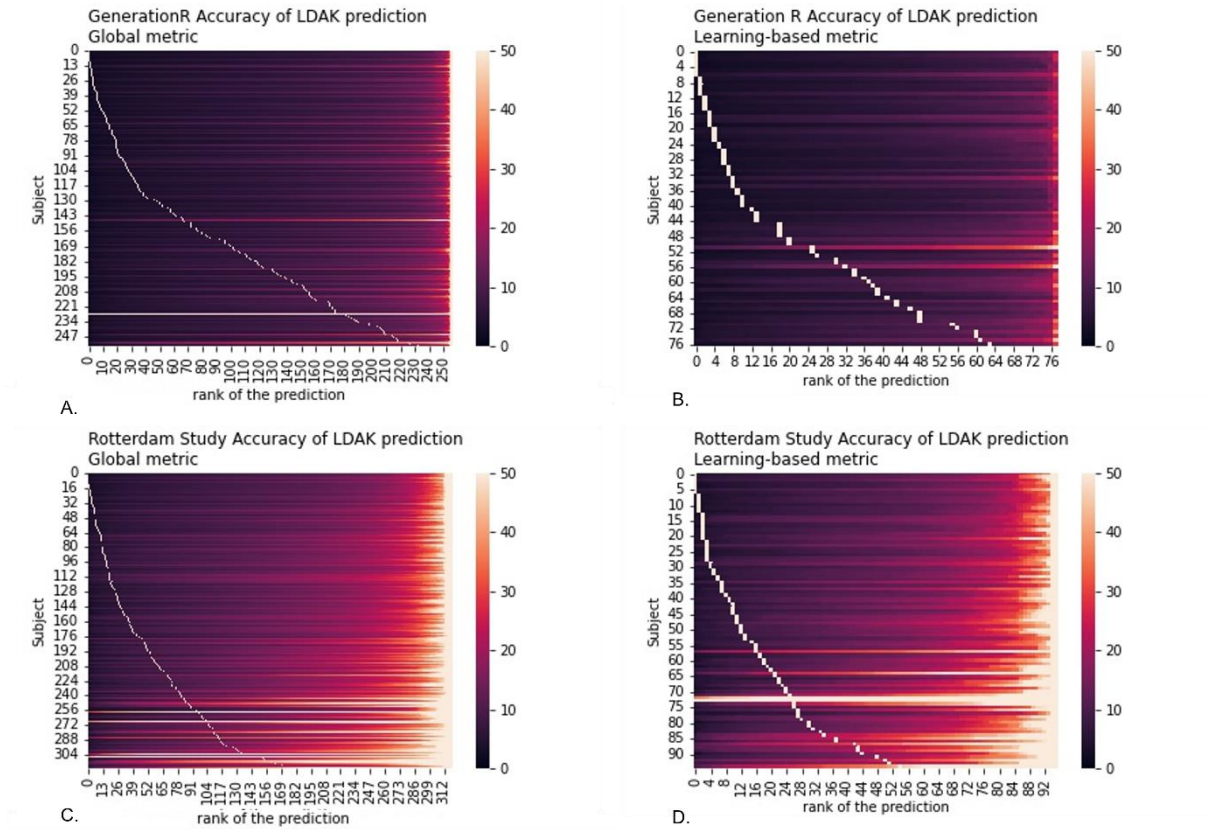

Figure 18: Heatmap plots of the LDAK prediction for the Generation R and the Rotterdam Study datasets using the second prediction evaluation measure. The top row (A-B) are the heatmaps for the Generation R dataset, figure A is for the global metric and figure B is for the le learning-based metric, which is why the x-axis has a smaller range. The bottom row (C-D) contains the heatmaps for the Rotterdam Study dataset. Figure C contains the result for the global similarity metric and D for the learning-based metric.

The datasets were also divided into sub-sets based on sex and ethnicity (for Generation R), and the similarity metrics and performance evaluation were computed. These results can be found in Figure 19.

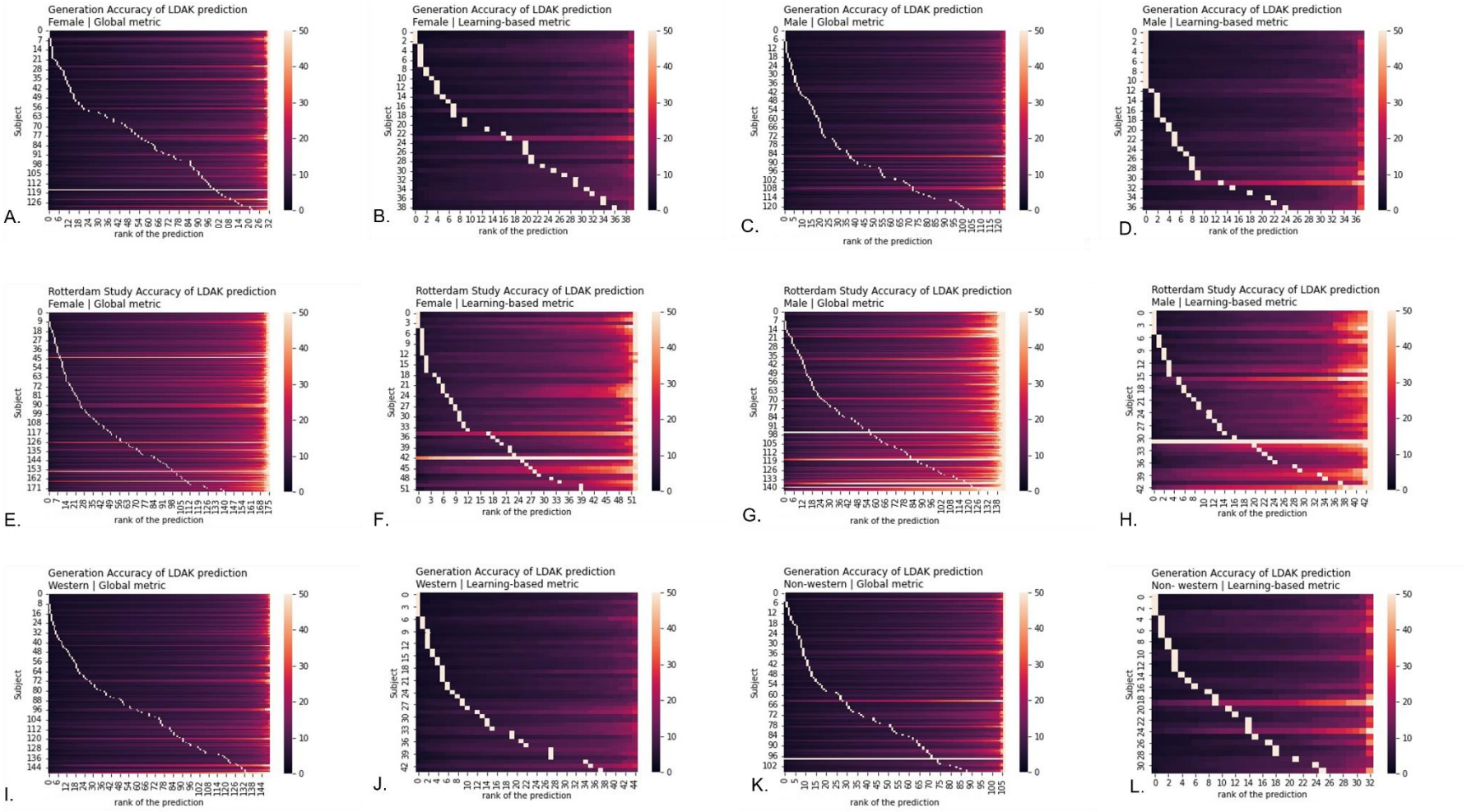

Figure 19: Heatmap plots for the different sub-sets of the Generation R and Rotterdam Study test sets. For each sub-set, both the global metric as well as the learning-based metric were computed. The top row (A-D) are the heatmap plots for the Generation R test set split by sex. The middle row (E – H) are the heatmap plots for the sub-sets of the Rotterdam Study. The bottom row (I – L) contains the heatmap plots for the Generation R data split by Ethnicity.

In addition to the heatmap plots, the AR was calculated for the second performance evaluation method. The AR was computed for each sub-set as well. The AR for the different datasets and sub-sets can be found in Table 4.

BLD-LDAK heritability model

	Global Metric		**Learning-based metric**	
	Generation R	*Rotterdam Study*	*Generation R*	*Rotterdam Study*
All	0,42	0,68	0,48	0,66
Sex				
Male	0,52	0,4	0,64	0,42
Female	0,3	0,54	0,26	0,52
Ethnicity				
Western	0,38		0,5	
non-western	0,46		0,42	

Table 4: Accuracy Ratio for each sub-set for the second predication evaluation measure.

5 Discussion

5.1 Interpretation of the results

The objective of this project was to explore the possibility of predictive modelling of facial features from DNA. Research on predictive modelling of facial features is in the early stage and has gotten more attention in recent years. In this project a genetic prediction model based on multilinear regression was used, called the LDAK-prediction model. In this chapter I will discuss the results found in the project and review the limitations of the project. Finally, recommendations for future research regarding this subject are described.

5.1.1 Reconstruction Error

The mean error found for the reconstruction of the endophenotypes into a facial mesh was in accordance to the reconstruction error described by the developers of the auto-encoder framework (37). In this project a mean error of 0,43 ± 0,08 mm was found, the mean error found by the developers was 0,426 ± 0,538 mm. The mean error for each vertex was visualized in Figure 11, this showed there were no regions with notable higher or lower mean errors. When looking at the histogram of all errors in Figure 12, it is visible that there are a small number of errors larger than one millimetre. One explanation for these outliers is noise in the ground truth due to facial hair such as the eyelashes and eyebrows. As visible in Figure 11, the eyes have a slightly higher mean error compared to the rest of the face, this is caused by the eyelashes. The eyelashes cause a rough surface around the eyes, which leads to a higher error between the ground truth and the reconstruction, since the latter one does not reconstruct the eyelashes.

The overall reconstruction error found is small and conform the error described in literature, with the highest error found around the eyes. As a result, the predicted faces and ground truths can be compared fairly, as the reconstruction error will not have large effect on the similarity measure. Thus, the similarity measure will be the result of inaccuracy of the prediction and not the reconstruction error.

5.1.2 Genetic prediction model

For the Generation R dataset and Rotterdam Study dataset the LDAK model was separately trained and tested. The R^2 found for the endophenotypes in the Generation R dataset were higher than those for the Rotterdam Study dataset. However, for both datasets the majority of the R^2 were negative. This indicates that the LDAK model was not well suited to model the endophenotypes based on the predictors (i.e. the SNPs). There are several explanations for this finding. First, the effect size of each SNP was computed based on the training samples. The LDAK documentation advices a sample sizes larger than 5000 subjects to compute the SNP effect size, unfortunately this number of samples was not available in the datasets used. Second, the LDAK model is restricted to linear relationships while it is unknown whether the relationship between facial morphology and genetics is linear. This will be discusses in more details in the limitations section. Third, facial morphology is a very complex, multidimensional trait which complicates the predictability and thus the performance of the genetic prediction model. The LDAK model is currently the best genetic prediction model for phenotype prediction, however it was not well suited for the complexity of genetic prediction of facial features.

5.1.3 First performance evaluation measure

Evaluating the performance of the genetic prediction model comes with several difficulties and there are several ways to analyse the performance of the prediction model depending on the objective. The complexity of the facial morphology poses a challenge for computing a similarity measure between

the predicted face and the ground truth. Since the face is a multipartite trait and consists of curves, surfaces and edges, it is difficult to capture all these features into one similarity measure.

When comparing at the two different metrics to measure the similarity between the predictions and the ground truths, there was not a larger difference found in the AR between the measures. It is noticeable however, that for the Generation R dataset, the results are better for the Global metric and for the Rotterdam Study the learning-based metric led to (slightly) higher AR's. It is important to highlight the difference in AR's are minimal and they may be the result of noise. Initially it was assumed the learning-based metric would lead to an improvement in the AR, since the 'important' vertices were given a higher weight. One explanation to why this expectation was not realised is due to the small sample sets available to train the weights. The test sets consisted only of 10% of the total datasets and from the test sets 70% was used for training the weights. Especially in the subsets (sex, ethnicity) these numbers were really low.

The results for the Generation R and the Rotterdam Study datasets are both for the global metric as the learning-based metric, apart from one instance, slightly better. This is in correspondence with the expectation that due to long-term exposure of environmental factors, the face gets less predictable from DNA over the years. However, only a small difference between the two datasets was found, where a larger difference was expected between the two datasets.

When splitting the datasets into subsets based on sex or ethnicity, no distinct change was found in the AR. In the LDAK-prediction model sex was used as an covariate, thus it was expected the performance of the prediction would not have increased for these subsets.

Finally, the AR's found were all only slightly above zero. An AR above zero indicates a predictive performance better than the baseline. The AR's found in this research were above this threshold, however only marginally. Therefor it is difficult to conclude whether the findings are the result of some predictive power of the genetic prediction model or the result of noise.

5.1.4 Second performance evaluation metric

In addition to the first performance evaluation metric a second metric was computed. The aim of this second metric was to investigate whether there was resemblance between individuals in the dataset as this could influence the evaluation of the prediction. For example, if two individuals, A and B, look similar to each other it can be expected that the predicted face of individual A has a high similarity to the ground truth of individual B. This effect could influence the outcome of the first performance evaluation metric. As visible in Figure 18 the similarity measure between the ground truths was quite high compared to the similarity measure between the predictions and ground truths. Which indicates there is a level of resemblance amongst individuals in the datasets. This is disadvantageous to the AR of the first performance evaluation metric. In addition, it is visible there are a few outliers in the dataset, such as subject 228 from Generation R and 260 for the Rotterdam Study, these subjects had a very low similarity measure to all other subjects in the test sets.

Similar to the first performance evaluation metric, there was no clear improvement in the AR's for the learning-based metric compared to the global metric. This is most likely due to the same cause: a low number of samples to train the weights.

The accuracy ratio's for the second performance evaluation metric are higher than those found for the first metric. The AR for the second metric indicates a fairly good ability to identify the predicted face corresponding to the ground truth from set containing other ground truths. However, the second performance evaluation metric is not suitable for determining the predictive performance of the genetic prediction model. The objective of the model is to predict the face as similar to the ground

truth face as possible. Therefore the first predication evaluation metric is the most important for the prediction performance.

5.2 Limitations

In addition to the findings of this thesis project, I would like to discuss a number of limitations of the project. First, the prediction model used in this project poses several limitations. The LDAK-model uses multilinear regression which is only able to model linear relationships between the genotype and the phenotype and restricts the prediction in that way. The exact effect of genes on facial morphology is unknown and restricting this relationship to linearity confines the research. The relationship could be non-linear due to gene-interactions, epistaxis or environmental contributors. In section 5.3 I will describe a method to model non-linear genotype-phenotype relations (38). With the LDAK-model each endophenotype is predicted separately, i.e. multilinear regression is performed for all two hundred independently. In this way the endophenotypes are assumed to be unrelated and have no influence on each other. However, research have found that different regions of the face are associated with the same genes (1). This suggests there is a connection between regions and modelling the endophenotypes independently does not allow this.

Second, the complexity of the facial morphology poses a big challenge for the performance evaluation of the prediction. Since the face consists of several regions and features, such as the nose, mouth and cheekbones, it is difficult to grasp this into one measure. In this project, the error for each vertex was computed and then the MSE of all 5023 vertex errors was calculated. This is limited as the error is the Euclidean distance thus there is no information on the direction of the error: the negative or positive direction in each plane. Moreover, with this method there is no evaluation of differences curves and surfaces.

Third, it is known that some regions are shaped predominantly by genes and others by the environment (1). Thus, it can be expected that the prediction of the regions influenced most by environment are less predictable and vice versa. With the learning-based metric it was attempted to account for this aspect and focus on the more predictable regions. However, we found no clear improvement in the results. This could be due to the small sample size that did not allow adequate training of the weights. A different method to encompass for the higher relative contribution of genes to some regions could be by predefining the weights for these regions, as it is known which regions are more susceptible to genetic influence.

Fourth, the developers of the LDAK software advise a sample set of at least 5000 subjects to perform a heritability analysis. For both datasets, this number of subjects was not available. Therefore, the heritability analysis is below optimal, which can result in incorrect effect sizes for the SNPs used for the genetic prediction model. To ensure an optimal heritability computation, more subjects should be included in the datasets.

Fifth, the reconstruction of the 3D facial mesh from the endophenotypes is restricted by the variation in faces in the training set. The auto-encoder learns the optimal reconstruction parameters based on a loss function, the error between the input and the output. When subjects differ from the subjects in the training data, the auto-encoder will not be able to reconstruct these faces correctly since it never had to handle the parameters before.

Sixth, from the Generation R dataset the facial images of children at the age of 9 years were used. It was assumed that the younger the subject, the more the face was shaped by genetics and less by

environmental factors, as there had been less exposure to these. However, this is not certain. For future research, facial images at different ages can be analysed.

Finally, the facial meshes were computed from images captured with a 3D camara system. Since, 3D facial meshes were used, the effect of head pose, and lighting are diminished. In addition, subjects were asked to maintain a neutral facial expression to reduce the influence of facial expression. However, a neutral expression in all subjects cannot be guaranteed, which may influence the prediction evaluation between the predicted and ground truth face. In addition, the 3D camera system captures the facial hear and skin texture, these are both absent in the predicted faces. Which also influences the prediction evaluation.

5.3 Future research

Predictive modelling of facial features from DNA is still in early stages and the objective of this project was to explore the topic and the possibilities. Future research should be conducted within this field and based on the findings in this project and the limitations, a few recommendations can be made. First, in this project a model based on multilinear regression was used, which restricts the prediction to linear relations between the genotype and phenotype. Non-linear prediction models should be explored f as they enable different relations between the genotype and phenotype. The GenNet framework is a possibility for a non-linear prediction model (39). It is a deep learning framework developed for phenotype predictions and enables the user to define their own architecture. In addition, the framework is designed to be interpretable and provides information on the weight assigned to each input, i.e. their contribution to the output.

Second, it is recommended to investigate different similarity metrics to evaluate the prediction accuracy. Due to the complexity of facial morphology, computing a meaningful similarity measure is challenging. However, improvements can be made, and more advanced methods should be explored. For example, focussing more on the regions that are associated with higher heritability. Or evaluation of the direction of the displacement in the x-, y- and z-plane, whether there are shifts in the negative or positive directions. An other possibility could be by performing a principal component analysis and analysing the principal components.

Third, larger sample sets are suggested for future research. To compute the heritability of the SNPs, more samples lead to better calculation of the effect size. And currently, the sample sets are smaller than advised. In addition, larger sample sets improve the power of the project. Since the face is high dimensional, more subjects available is better for the prediction. Moreover, more variation in ethnicity and age is recommended. As ancestry and facial morphology are highly related (1).

Fourth, currently only a small number of SNPs associated with the face have been identified. For that reason, in this project all SNPs available were used for the prediction model. In the future, when more SNPs have been identified, the genetic prediction model could be based on these SNPs instead of all SNPs available.

Finally, some regions of the face are predominantly influenced by genes and others by environmental factors (1). In future research, the focus could be on the regions mainly associated with genetics. For example by predefining weights for area's that are associated with a higher genetic contribution. Genetic prediction of the regions shaped predominantly by environmental factors will not be successful, as you are trying to model a relationship that does not exist.

6 Conclusion

The objective of this study was to explore the possibilities of predictive modelling of facial features from DNA. The results found in this study indicate the presence of (some) predictive power, however this is currently very limited. The results for the Generation R dataset were slightly higher than for the Rotterdam Study dataset. Due to long-term exposure to environmental factors the contribution of the environmental component to the face increases with age, thus prediction based on genetics becomes less accurate. This assumption seemed somewhat true, however no substantial difference in prediction accuracy was found.

Predictive modelling of the facial morphology is a complex task due to the high-dimensionality of the face and the complexity of a suitable similarity measure, amongst other things. Research in this field is still in an early stage and although the predictive power of the model is currently low, the results warrant further research. Future research could be in the direction of different prediction models that allow modelling of non-linear relationships, improved metrics for computing the similarity measure between two faces and predictions based solely on genetic loci associated with the face.

7 Ethical Analysis

Technology has greatly impacted medical practice and continues to do so with the continuous development of new innovations. With the increasing role of technology in healthcare, new ethical questions arise. My master program focusses on the contributions of medical technology, on how it aides in therapies and the value of technology in healthcare, but there is little attention to the ethical side of medical technology. During my internships I saw some of the ethical dilemmas that arise as a result of medical technology and I wanted to know more about how to analyse and carry out a careful evaluation of these kind of questions. For that reason I wanted to do an internship at the department of Medical Ethics, Philosophy and History of Medicine. During this internship I had the chance to learn more about the ethics of medicine and medical technology, and got to participate in a research project about the ethics of eHealth technology. For my master thesis I wanted to include this side of medical technology, therefore I carried out an ethical analysis of my thesis subject. In this chapter I will discuss the analysis and give a background to the relevance of ethics of medical technology.

7.1 Introduction

Ethics has been connected to medicine since Antiquity, it was Hippocrates who first linked ethical principles to the responsibilities of a doctor. Being a competent doctor meant more than being skilled in medical procedures, it encompasses taking account of your moral obligations as a professional (40). Nowadays, ethical committees are in place to evaluate moral questions, collaborating with medical professionals, patients and other relevant parties to evaluate the ethical point of view. With the rapid advancement of technology in medicine, new moral questions arise continuously. The first time technology was used in medicine, was in 1816 when the stethoscope was introduced by French physician René Laennec (41). Since then technology has rapidly expanded into all areas of medicine, irreversibly changing the relationship between patient and doctor, but also the way we define sickness and health (42,43). Technology has provide life-sustaining treatments, can take over organ functions or visualize the insides of the human body without damaging its integrity. In recent years, medical technology has moved towards the field of data science (44). Enormous amounts of data are gathered through electronic medical records, which offers new opportunities. Data science and artificial intelligence offer possibilities such as identifying risk factors and analysing vital signs (43,45). Technology has enlarged the power of medicine, but also led to new ethical questions in the field of medical practice. Careful evaluation of these questions is necessary to protect patients autonomy and privacy (46). Moreover, exploring ethical issues can be an opportunity for enhancing technology by starting a debate and requiring an analysis of different standpoints (47).

There are several methods to assess the ethical dimensions of medical technology. For the analysis of my thesis subject I have chosen to use the framework created by Prof. Dr. Schermer, Chair of the department Medical ethics, philosophy and history of medicine at Erasmus MC (48). The framework can be used to systematically map the ethical aspects of technology. The framework consists of four steps: goals, means, unintended effects and finally an evaluation (48). First, the goals of the technology are analysed, and whether these are at the benefit of the subject or for other stakeholders. Other stakeholders could be medical professionals, family members or insurance companies. Second, the means of the technology are evaluated (48). This translates in an analyses of the intention of the technology, is this technology morally acceptable. In addition, the risks and the secondary effects of the technology are analysed. Third, the unintentional effects are explored. The effect of medical technology is often not limited to their intended purpose, there are often secondary effects to new technologies (40,48). Since these are unexpected effects it is difficult to fully describe them, however an educated guess can be made. And the moral acceptability of these effects is evaluated. Finally, an assessment of is made whether the technology is proportional to the goal and whether the pro's and

con's are in harmony. It is important to take the goals, pro's and cons and the unintentional effects all into account (48).

1. Goals

In this thesis project, the medical technology explored was predictive modelling of facial features from DNA using a multilinear regression model. The goal of predictive modelling of facial features is to get an understanding of the genetic architecture underlying facial features and the effect of genetics on the phenotype. In the past, research regarding facial genetics was focussed on the genetics of diseases and anomalies, and not so much on the genetics of normal-range facial variation (1). In recent years, genes associated with facial morphology have been identified, however their effect on the 'phenotype' is not known. Insight in these effects could aide in comprehending the genetic basis of disease and unravelling the contributions of genetic and environmental factors on the face. The stakeholders in this goal are researchers and medical professionals, as it would lead to a bigger understanding of facial genetics. The technology will not lead to individual benefits for the subjects in the dataset used during this project. At this stage it is not possible to say the goal is met, as the research is still at a very early stage and it is an explorative project. Moreover, there is a chance that the goal mentioned will never be met, since it is a very complex task which may even not be possible.

2. Means

I believe the intention of the technology itself is good. The technology used is a multilinear regression model, which is a morally acceptable technology, it poses no harm or risk to users. The risks and secondary effects of the technology are, I believe, mainly in the data used for the model and the interpretation of the results. The data used for the model has a large impact on the reliability of the results as it could lead to biases or a distorted view. A model is only capable of learning relationships that are present within the input data, therefore it is limited by the diversity in this set. For example, if an input set solely contains male subjects, the model is not have any predictive power for female subjects. Or when there is no ethnic diversity in the input dataset, the model is only useful for that particular ethnicity. It is very important to be aware of these implications and consider them carefully. In addition, the results can be processed in different ways which influences the outcome and conclusions made. Close examination of the results and evaluation methods is crucial to draw reliable and fair conclusions.

3. Unintended effects

As mentioned earlier, it is difficult to describe the unintended effects. However, when brainstorming about the possible unintended effects the following things came to mind. First, promising preliminary results could lead to an idea of in depth understanding of the genetic architecture of facial morphology. This could hold back future research and lack of a critical view on the subject, where you are not open to exceptions or oddities. Second, this technology could lead to the idea of being able to fully predict the face from DNA. This notion is untrue, since the face is not solely shaped by genetics. Facial morphology is the result of genetic, environmental and epigenetic factors and their interactions (1). Thus, full prediction of the face from DNA is not possible. It is important to emphasize this and keep this in mind. When this is disregarded and the belief exists the prediction is perfect, it could lead to illegitimate conclusions. Especially in the field of forensics, where this could for example potentially lead to wrongful identification and accusation.

4. Assessment

When evaluating whether the technology is proportional to the goal, weighing the pro's, cons and unintended effects, I believe that this technology is acceptable. Currently, there are no other methods available to investigate the goal of this technology. In addition, little is known about the magnitude of

influence of the genotype on normal-range variation in facial morphology. Insights in the genetic architecture could aide in, amongst other thigs, understanding the genetic basis of diseases and the relative contribution of genotype on facial morphology. Currently, this research is still in a very early stage and the feasibility of the technology remains a question. If research on this topic continues to develop and proves to be feasible, there are other fields of study that see potential in this technology, such as anthropology and forensics. The applications in these fields of study come with their own ethical questions and should be carefully evaluated if research on the topic moves in that direction. Especially applications of this technology for forensic purposes raise ethical questions (regarding autonomy, privacy and justification) and application in this field should be carefully evaluated as it could lead to worrisome situations, such as wrongful identifications or accusations. Due to the extensiveness of this particular application a thorough ethical analysis is required, for example using the Four Principles Approach by Beauchamp and Childress (49)

8 References

1. Richmond S, Howe LJ, Lewis S, Stergiakouli E, Zhurov A. Facial genetics: A brief overview [Internet]. Vol. 9, Frontiers in Genetics. Frontiers Media S.A.; 2018. p. 462. Available from: www.frontiersin.org

2. Djordjevic J, Zhurov AI, Richmond S. Genetic and environmental contributions to facial morphological variation: A 3D population-based twin study. PLoS One. 2016;

3. Fagertun J, Wolffhechel K, Pers TH, Nielsen HB, Gudbjartsson D, Stefansson H, et al. Predicting facial characteristics from complex polygenic variations. Forensic Sci Int Genet. 2015 Nov 9;19:263–8.

4. Claes P, Liberton DK, Daniels K, Rosana KM, Quillen EE, Pearson LN, et al. Modeling 3D Facial Shape from DNA. Luquetti D, editor. PLoS Genet [Internet]. 2014 Mar 20;10(3):e1004224. Available from: https://dx.plos.org/10.1371/journal.pgen.1004224

5. Liu X, Kayser M, Kushner SA, Tiemeier H, Rivadeneira F, Jaddoe VW V., et al. Association between prenatal alcohol exposure and children's facial shape. A prospective population-based cohort study. medRxiv [Internet]. 2021 Jul 25;2021.07.22.21260946. Available from: https://www.medrxiv.org/content/10.1101/2021.07.22.21260946v1

6. Speed D, Cai N, Consortium U, Johnson MR, Nejentsev S, Balding DJ. Re-evaluation of SNP heritability in complex human traits Europe PMC Funders Group. Nat Genet [Internet]. 2017;49(7):986–92. Available from: http://www.nature.com/authors/editorial_policies/license.html#terms

7. Cole JB, Manyama M, Larson JR, Liberton DK, Ferrara TM, Riccardi SL, et al. Human facial shape and size heritability and genetic correlations. Genetics [Internet]. 2017 Feb 1;205(2):967–78. Available from: https://pubmed.ncbi.nlm.nih.gov/27974501/

8. Roosenboom J, Hens G, Mattern BC, Shriver MD, Claes P. Exploring the Underlying Genetics of Craniofacial Morphology through Various Sources of Knowledge. Biomed Res Int [Internet]. 2016;2016:3054578. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28053980

9. Claes P, Roosenboom J, White JD, Swigut T, Sero D, Li J, et al. Genome-wide mapping of global-to-local genetic effects on human facial shape. Nat Genet [Internet]. 2018 Mar 1 ;50(3):414–23. Available from: https://doi.org/10.1038/s41588-018-0057-4

10. White JD, Indencleef K, Naqvi S, Eller RJ, Hoskens H, Roosenboom J, et al. Insights into the genetic architecture of the human face. Nat Genet [Internet]. 2021 Jan 1 ;53(1):45–53. Available from: https://doi.org/10.1038/s41588-020-00741-7

11. Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacón-Duque JC, et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nat Commun [Internet]. 2016 Mar 1;7(1):1–12. Available from: www.nature.com/naturecommunications

12. Paternoster L, Zhurov AI, Toma AM, Kemp JP, St. Pourcain B, Timpson NJ, et al. Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. Am J Hum Genet [Internet]. 2012 Mar 9;90(3):478–85. Available from: https://pubmed.ncbi.nlm.nih.gov/22341974/

13. Liu F, van der Lijn F, Schurmann C, Zhu G, Chakravarty MM, Hysi PG, et al. A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. Gibson G, editor. PLoS Genet [Internet]. 2012 Sep 13;8(9):e1002932. Available from: https://dx.plos.org/10.1371/journal.pgen.1002932

14. Cole JB, Spritz RA. The Genetics of Facial Morphology. eLS [Internet]. 2017 Sep 15;1–9. Available from: http://doi.wiley.com/10.1002/9780470015902.a0027240

15. Shaffer JR, Orlova E, Lee MK, Leslie EJ, Raffensperger ZD, Heike CL, et al. Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. Barsh GS, editor. PLOS Genet [Internet]. 2016 Aug 25;12(8):e1006149. Available from: https://dx.plos.org/10.1371/journal.pgen.1006149

16. Cha S, Lim JE, Park AY, Do JH, Lee SW, Shin C, et al. Identification of five novel genetic loci related to facial morphology by genome-wide association studies. BMC Genomics [Internet]. 2018 Jun 19;19(1):481. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4865-9

17. Crouch DJM, Winney B, Koppen WP, Christmas WJ, Hutnik K, Day T, et al. Genetics of the human face: Identification of large-effect single gene variants. Proc Natl Acad Sci U S A [Internet]. 2018 Jan 23;115(4):E676–85. Available from: https://pubmed.ncbi.nlm.nih.gov/29301965/

18. Lee MK, Shaffer JR, Leslie EJ, Orlova E, Carlson JC, Feingold E, et al. Genome-wide association study of facial morphology reveals novel associations with FREM1 and PARK2. Li Y, editor. PLoS One [Internet]. 2017 Apr 25;12(4):e0176566. Available from: https://dx.plos.org/10.1371/journal.pone.0176566

19. Xiong Z, Dankova G, Howe LJ, Lee MK, Hysi PG, De Jong MA, et al. Novel genetic loci affecting facial shape variation in humans. Elife. 2019 Nov 1;8.

20. Walsh S. DNA Phenotyping: The Prediction of Human Pigmentation Traits from Genetic Data [Internet]. 2013. Available from: http://www.rpi.edu/dept/NewsComm/sub/photos/dna_eye.jpg

21. Chaitanya L, Breslin K, Zuñiga S, Wirken L, Pośpiech E, Kukla-Bartoszek M, et al. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. Forensic Sci Int Genet. 2018 Jul 1;35:123–35.

22. Fitzgerald R, Graivier M. Update on Facial Aging. Aesthetic Surg J [Internet]. 201AD;30(1):11S-24S. Available from: http://www.sagepub.com/www.aestheticsurgeryjournal.com

23. Chen L. Curse of Dimensionality. In: Encyclopedia of Database Systems [Internet]. Springer, Boston, MA; 2009. p. 545–6. Available from: https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_133

24. Zhang Q, Privé F, Vilhjálmsson B, Speed D. Improved genetic prediction of complex traits from individual-level data or summary statistics. Available from: https://doi.org/10.1038/s41467-021-24485-y

25. Het onderzoek - Generation R [Internet]. Available from: https://generationr.nl/algemeen/het-onderzoek/

26. Dept. of Epidemiology [Internet]. Available from: http://www.epib.nl/research/ergo.htm

27. Kooijman MN, Kruithof CJ, van Duijn CM, Duijts L, Franco OH, van IJzendoorn MH, et al. The Generation R Study: design and cohort update 2017. Eur J Epidemiol. 2016;31:1243–64.

28. Liu X, Kayser M, Kushner SA, Tiemeier H, Rivadeneira F, Jaddoe VW V, et al. Association between prenatal alcohol exposure and children ' s facial shape . A prospective population-based cohort study. Supplementary. :1–13.

29. Arfan Ikram M, Guy Brusselle ·, Ghanbari M, Goedegebure · André, Kamran Ikram · M,

Kavousi M, et al. Objectives, design and main findings until 2020 from the Rotterdam Study. Eur J Epidemiol [Internet]. 2020;35:483–517. Available from: https://doi.org/10.1007/s10654-020-00640-5

30. Kruithof CJ, Kooijman MN, van Duijn CM, Franco OH, de Jongste JC, Klaver CCW, et al. The Generation R Study: Biobank update 2015. Eur J Epidemiol 2014 2912 [Internet]. 2014 Dec 21 ;29(12):911–27. Available from: https://link.springer.com/article/10.1007/s10654-014-9980-6

31. Ikram MA, Brusselle GGO, Murad SD, Duijn CM van, Franco OH, Goedegebure A, et al. The Rotterdam Study: 2018 update on objectives, design and main results. Eur J Epidemiol [Internet]. 2017 Sep 1;32(9):807. Available from: /pmc/articles/PMC5662692/

32. Quality Control | dougspeed.com [Internet]. Available from: http://dougspeed.com/quality-control/

33. Goodfellow I, Bengio Y, Courville A. Deep Learning [Internet]. MIT Press; 206AD. 499–523 p. Available from: https://www.deeplearningbook.org/

34. Jordan J. Introduction to autoencoders. [Internet]. 2018. Available from: https://www.jeremyjordan.me/autoencoders/

35. Dertat A. Applied Deep Learning - Part 3: Autoencoders | by Arden Dertat | Towards Data Science [Internet]. 2017. Available from: https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

36. BLD-LDAK Annotations | dougspeed.com [Internet]. Available from: http://dougspeed.com/bldldak/

37. Gong S, Chen L, Bronstein M, Zafeiriou S. SpiralNet++: A Fast and Highly Efficient Mesh Convolution Operator 3 FaceSoft.io 4 Twitter.; Available from: https://github.com/sw-gong/spiralnet_plus

38. Sverdlov S, Thompson EA. The Epistasis Boundary: Linear vs. Nonlinear Genotype-Phenotype Relationships. bioRxiv [Internet]. 2018 Dec 21 [cited 2021 Sep 18];503466. Available from: https://www.biorxiv.org/content/10.1101/503466v1

39. van Hilten A, Kushner SA, Kayser M, Arfan Ikram M, Klaver CC, Niessen WJ, et al. GenNet framework: interpretable neural networks for phenotype prediction 2 3. [cited 2021 Jan 8]; Available from: https://doi.org/10.1101/2020.06.19.159152

40. Have ten H. Leerboek Ethiek in de gezondheidszorg. 2019. 19–36 p.

41. Reiser S. Revealing the body's whispers. How the stethoscope transformed medicine. In: Technological Medicine: The Changing World of Doctors and Patients. 2009. p. 1–13.

42. Ng T. Ethics in the age of medical device technologies. Virtual Mentor. 2007.

43. Ali YR, Rizi Hossein Ali Y. Ethical impact of the technology on the healthcare system. Res Artic J Clin Investig Stud J Clin Invest Stud. :2020.

44. Arboleda P. As devices generate more data, AI is becoming indispensable for medtech [Internet]. 2019. Available from: https://ncube.com/blog/data-science-in-healthcare-10-ways-of-industry-transformation

45. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine [Internet]. Vol. 315, JAMA - Journal of the American Medical Association. American Medical Association; 2016 [cited 2021 Jun 21]. p. 551–2. Available from:

https://pubmed.ncbi.nlm.nih.gov/26864406/

46. Rigby MJ. FROM THE EDITOR Ethical Dimensions of Using Artificial Intelligence in Health Care. AMA J Ethics [Internet]. 2019 [cited 2021 Aug 25];21(2):121–4. Available from: www.amajournalofethics.org

47. Toom V, Wienroth M, M'charek A. Correspondence. Forensic Sci Int Genet [Internet]. 2016;22. Available from: http://dx.doi.org/10.1016/j.fsigen.2016.01.010

48. Schermer M. Gedraag je! Ethische aspecten van gedragsbeïnvloeding doornieuwe technologie in de gezondheidszorg [Internet]. Rotterdam; 2007. Available from: https://www.academia.edu/469666/Gedraag_je_Ethische_aspecten_van_gedragsbeinvloeding_door_nieuwe_technologie_in_de_gezondheidszorg

49. Beaucamp TL, Childress JF. Principles of Biomedical Ethics [Internet]. 5th ed. Oxford University Press; 2001. 454 p. Available from: https://books.google.nl/books?hl=nl&lr=&id=_14H7MOw1o4C&oi=fnd&pg=PR9&ots=1xUh3LAkSt&sig=mNnrQTgnk3UDUWwowWPL1jYGzF4&redir_esc=y#v=onepage&q&f=false