Age and sex-appropriate 3D shape analysis and geometric deep learning for children undergoing craniofacial diagnostics and surgery

A Data Management Plan created using DMPonline.be

Creator: Peter Claes

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: G0D1923N

ID: 197904

Start date: 31-12-2022

End date: 30-12-2026

Project abstract:

Facial growth is of interest in biomedicine and computer science. In biomedicine, assessing facial abnormality in children is crucial to optimal care in both clinical genetics and craniofacial surgery, yet is complicated by the normal growth of the face. This complication was traditionally resolved using age and sex-specific standards of craniofacial measurements (growth curves) that oversimplify facial shape. More advanced techniques using 3D shape data are clearly needed. From a computer science point of view modelling facial growth from image data is an ideal challenge for new non-linear methodologies. Conditional generative modelling using deep learning shows promise in modelling growth and ageing in 2D image data, but three-dimensional methods are needed to unlock their biomedical potential. In a crucial development we recently released the first generation of ‘3D Growth curves’ based on conditional generative modelling using multivariate shape analysis rather than deep learning. First, we deploy these 1st generation growth curves towards medical assessments in clinical genetics and craniofacial surgery to understand their potential. In parallel, we develop 2nd generation growth curves, extending the state of the art in deep conditional generative models towards anatomical 3D shape data. Finally, both generations are evaluated from both biomedical and computational perspectives.

Last modified: 08-06-2023

Age and sex-appropriate 3D shape analysis and geometric deep learning for children undergoing craniofacial diagnostics and surgery
FWO DMP (Flemish Standard DMP)

---

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

			Only for digital data	Only for digital data	Only for digital data
Dataset Name	Description	New or reused	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)
AHEAD	The Australian Head Examination and Assessment Dataset (n=853), 3D facial images DOI: 10.1038/s41598-018-22752-5, collected by Melbourne Childrens' Campus (MCC) Melbourne Australia	Reuse	Observational	.obj wavefront	<100GB
AHEAD meta	Demographic metadataof the AHEAD dataset	Reuse	Observational	.xlsx excel	<100MB
3DFN	The Facebase normative dataset (n=1886), 3D facial images (https://www.facebase.org/facial_norms/)	Reuse	Observational	.obj wavefront	<100GB
3DFN meta	Demographic metadata of the 3DFN dataset	Reuse	Observational	.xlsx excel	<100MB
ADAPT	ADAPT – Anthropology, DNA, and the Appearance and Perception of Traits Study, PennSate, US (n=1880) , 3D facial images https://ched.la.psu.edu/projects/adapt-anthropology-dna-and-the-appearance-and-perception-of-traits-study-ongoing/	Reuse	Observational	.obj wavefront	<100GB
ADAPT meta	Demographic metadata of the ADAPT dataset	Reuse	Observational	.xlsx excel	<100MB
ALSPAC	The Avon Longitudinal Study of Parents and Children, (n~3000), 3D facial images, http://www.bristol.ac.uk/alspac/	Reuse	Observational	.obj wavefront	<100GB
ALSPAC meta	Demographic metadata of the ALSPAC dataset	Reuse	Observational	.xlsx excel	<100MB
ABCD	The longitudinal Adolescent Brain and Cognitive Development (ABCD) study, full head MRI images (n~12.000), https://abcdstudy.org/	Reuse	Observational	.nifti	<50TB
ABCD meta	Demographic and imaging metadata of the ABCD dataset	Reuse	Observational	.xlsx excel	<100MB
SYND	Developing 3D Craniofacial Morphometry Data and Tools to transform Dysmorphology (http://doi.org/10.25550/1WWC, hosted at www.FaceBase.org, 3D facial images (n~3000)	Reuse	Observational	.obj wavefront	<100GB
SYND meta	Demographic and clinical metadata of the SYND dataset	Reuse	Observational	.xlsx excel	<100MB
MeshMonk	Non-rigid surface registration toolbox, Opensource software toolbox for mapping 3D facial images. https://doi.org/10.1038/s41598-019-42533-y; Medical Imaging Research Center / Mesh Monk · GitLab (kuleuven.be)	Reuse	Software	C++, .py, .m	<10MB
MRI2Face	In-house software pipeline for the robust extraction of 3D facial surfaces from full head 3D MRI T1 Scans, used for the processing of the ABCD dataset from MRI images to 3D surface data as explained here: https://doi.org/10.1101/2022.12.01.518684	Reuse	Software	.py	<10MB
MV-3DGrowth	Software scripts to 1) model a 3D facial growth curve from cross sectional 3D facial image datasets, 2) assess a 3D facial image using an age and sex appropriate reference using Multivariate Techniques.https://doi.org/10.1038/s41598-021-91465-z; https://github.com/harrymatthews50/3DGrowthCurves	Reuse	Software	.m	<100MB
DL-3DGrowth	To develop scripts to 1) model a 3D facial growth curve from cross sectional 3D facial image datasets, 2) assess a 3D facial image using an age and sex appropriate reference using geometric deep learning.	New development	Software	.m, .py	<100MB
Clinical Analysis	To develop scripts to run project specific clinical studies based on the available datasets and software tools.	New development	Software	.m(at), .r, .py(w),	<100MB

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

AHEAD: DOI:10.1038/s41598-018-22752-5
3DFN: https://www.facebase.org/facial_norms/
ADAPT: https://ched.la.psu.edu/projects/adapt-anthropology-dna-and-the-appearance-and-perception-of-traits-study-ongoing/
ALSPAC: http://www.bristol.ac.uk/alspac/
ABCD: https://abcdstudy.org/
SYND: http://doi.org/10.25550/1WWC

MeshMonk:https://doi.org/10.1038/s41598-019-42533-y; Medical Imaging Research Center / Mesh Monk · GitLab (kuleuven.be)
MRI2Face: https://doi.org/10.1101/2022.12.01.518684
MV-3DGROWTH:https://doi.org/10.1038/s41598-021-91465-z; https://github.com/harrymatthews50/3DGrowthCurves

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

The overall aim of this project is to build and evaluate sex-specific 3D facial growth curves for the benefit of children undergoing craniofacial clinical diagnostics and surgical intervention. Therefore, we will be processing human subject data, more specifically 3D facial images, demographic data (sex & age) and clinical data (e.g., diagnosis of clinical condition and treatment plan). This is a retrospective study, where existing datasets (AHEAD, 3DFN, ADAPT, ALSPAC, ABCD, and SYND) are being processed into a data-anonymous open-access model. No new data is being collected within the scope of the project. Ethics approval: S60568, Statistical Genetic Analysis of Quantitative Facial Variation, S56392, Directional & Fluctuating decomposition and analysis of 3D Facial Discordancy: Insights into disordered growth).

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

We plan to use the following categories of personal data: Personal details such as sex and age, physical traits such as length and weight, 3D facial images not for the purpose of uniquely identifying a natural person, data revealing racial or ethnic origin (facial images and genetic ancestry), data from or linked to the Electronic Patient Record (EPR).
Pseudonymization: As many identifiers as possible have been removed from the data. The data was pseudonymized by the research partners and data repositories (AHEAD, 3DFN, ADAPT, ALSPAC, ABCD, and SYND) with whom the (DTA, MTA) agreements are in place. Facial images are used, hence full anonymization is not possible, but re-identification is prohibited as outlined in the agreements.
Data minimization: For all 3D facial images, shape information about the face is sufficient for the project and no texture (i.e. the recognizable facial photographs that are projected onto the unicolor shape) data are needed and were therefore removed, if not already done so by the data repository itself. This again greatly reduces the risk of recognition, since publicly available facial recognition software is not applicable. No individual pictures will be published.
The major outcome of the research are updated and open-access 3D facial growth curves. These curves present themselves as models, that can be used by other researchers or clinicians in the field, without the need to access the underlying data, and without the risk of re-identifying the underlying data.
An associated PRET application specific to this project has been submitted and approved in the meantime: G-2023-6623.

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

This is a retrospective study, combining multiple existing data sources, many of which (3DFN, ALSPAC, ABCD, SYND) are gated-access repositories with DTA and MTA agreements restricting commercial exploitation of the results. The DTA and MTA's for the ADAPT and AHEAD datasets also limit the use for commercial exploitation of the data. In part due to these restrictions, also due to the open-access ambitions of the project itself, resulting 3D facial growth curves are planned to become open-access.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

In collaboration with the DMP support team (Naeem Muhammad, Veerle Van den Eynden) at the KU Leuven, we are currently working on a protocol that will help individual researchers and overall project managers to organize and document, the data flow (from original data to processed data and results) and software development (from original concept to shareable and documented implementation). A standard folder structure per project/subproject

involves:
Data: separate folder for all fixed raw data used in the project that do not change throughout the project. In case the data is large and already exist on other dataserver partitions, explicit hyperlinks to the data will be used.
Preprocessing: Subfolders for each step in preprocessing.
Analysis: Subfolders are created for each event / step / pipeline.
Sourcecode: All models, code you will use for preprocessing and analysis.
Documentation: Any documentation files for the project. This can include papers.
Datapoints: Folder for intermediate datasets that you will use again in future.
Temp: Folder for draft versions of data and codes you want to keep.
The project folder includes a README file that lists the folder structure with a brief description of the content of each folder.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

For the image data the following metadata standards will be used. .obj wavefront files, is one of the most common standards to store and load 3D surfaces data (AHEAD, 3DFN, ADAPT, ALSPAC, and SYND). Nifti and Dicom files, are used for structural MRI T1 images (ABCD), the headers in these images also contain metadata on the subject, but more importantly, on the imaging protocol and device. The latter is important working with MRI images since, images and noise in the images are scanner and protocol dependent.
For the demographic metadata, we will work with excel workbook .xlsx and .csv files. We will impose the use of an image identifier that is the (anonymized) name of the image file to be linked to, followed by information on sex and age primarily.
For the genetic data (ABCF, 3DFN, ADAPT, ALSPAC), the following two most common metadata standards will be used, PLINK file format (.bim, .bed and .fam) and The Variant Call Format (.vcf). The VCF format reference specifies the format of a text file used in bioinformatics for storing gene sequence variations. The format has been developed with the advent of large-scale genotyping and DNA sequencing projects, such as the 1000 Genomes Project.
For the software data files, development thereof is mainly done in python and matlab. For both environments, but especially for python, version and library documentation will be added in the documentation and readme files stored in the same location and the source code.

3. Data storage & back-up during the research project

Where will the data be stored?

Supported by ESAT-PSI members located at the campus Gasthuisberg, we make use of a dedicated data server: HPE Apollo 4510 Gen10, with 58 disks of 16TB, (Total storage 760 Tb) with 2 RAID sets defined one for Data and one for Backup. This data server is located within the data center of the UZ Leuven Hospital.

How will the data be backed up?

A one-time copy of the data downloaded from the public depositories is stored in dedicated data partitions of the data server. No further backup of this data (considered as original data) is made, since they are re-downloadable from their respective repositories. Access to this one-time copy is read only and based on access rights for individual researchers when they are named explicitly on the DTA agreements.
For data processed, data points, and generated results, the data server is equipped with an incremental backup system. In addition, once per month, a backup is made on offline storage operated by 2 mac minis, each connected to three Promise Pegasus3 R8 systems.
For source code data, researchers are encouraged to use KU Leuven GitLab, providing incremental version control.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

Given the data size limits described above, the data server is large enough to handle project storage and backup.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The data server is physically located in the secured data center of the Hospital UZ Leuven. Therefore, the security is at the level of security needed to maintain and store sensitive patient information. Access is password protected, where it is mandatory to change the password every six months and following strict rules to ensure strong passwords only. For remote access from outside the hospital, a two-factor authentication is required. Inside the hospital, front end computers with access to the data server are within access-restricted premises with badge control. In addition to general password protection and two-factor authentication, the different data partitions and data folders are linked to a restricted user access and control management system. This allows us to dedicate access to individual researchers, based on their involvement and listing on the DTA agreements. Our IT manager, Dominique Delaere, maintains the data server and its access management system.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

With the data server a cost of 10.8€ for 1Tb/year is estimated. This was a one time cost, already covered by additional funding.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

The “original data” as downloaded from the data repositories will only be kept for the duration as stipulated in the respective DTA agreements. Typically, these agreements request the deletion of the data downloaded after finishing the project.
Any other data relevant as output from the project involving relevant datapoints and results needed to replicate the work, generated source code, and resulting open-access models will be retained for at least five years, or as long as needed by the requirements of the KU Leuven and FWO.

Where will these data be archived (stored and curated for the long-term)?

A separate data partition “RDM” (Research Data Management), was created on the data server to serve the purpose of data preservation. The data archiving is organized for each paper published, each PhD defended, and each project finished.
Open-access results from the project will also be uploaded as open-access on specific platforms (e.g., GitLab) and data repositories (I.e., the original data repository where the data was downloaded from, and where results are requested to be uploaded, e.g., ALSPAC).

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

With the data server a cost of 10.8€ for 1Tb/year is estimated. Renewal of the data server will be covered by additional funding and in collaboration with members of ESAT-PSI located at the campus Gasthuisberg.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Other, please specify:

There is no collection of new human subject data in this project. Data used in this project is already made available for reuse by gated-access data repositories controlled and managed by others.
Source code and 3D facial growth models will be made available.

If access is restricted, please specify who will be able to access the data and under what conditions.

Not applicable

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Not applicable for the data from repositories. No restrictions are expected for the source code and resulting 3D facial growth model.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Source code and 3D facial growth models will be made available using GitLab.

When will the data be made available?

When related papers are published.

Which data usage licenses are you going to provide? If none, please explain why.

None, for the open-access models and source code.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the

comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

No expected costs.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

(Meta)data will be documented by the researcher(s) working on the project at the time of analysis and compiled at regular intervals as well as when preparing for publication.

Who will manage data storage and backup during the research project?

The researcher(s) with the help of the IT-manager (Dominique Delaere) ensure secure and backed-up data storage.

Who will manage data preservation and sharing?

The project (co-)supervisors will manage the data preservation and sharing, with support from the individual researchers and the IT-manager.

Who will update and implement this DMP?

The project (co-)supervisors bear the final responsibility for all data management during and after data processing, including updating & implementing this DMP. The DMP will be evaluated at regular meetings between the researcher and the (co-)supervisors during the project.