Artificial intelligence for deep profiling and characterization of colorectal polyps

A Data Management Plan created using DMPonline.be

Creators: Alexander Jans, n.n. n.n., Raf Bisschops

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Raf Bisschops, Alexander Jans, n.n. n.n.

Data Manager: Raf Bisschops, Alexander Jans, n.n. n.n.

Project Administrator: Raf Bisschops, Alexander Jans, n.n. n.n.

Grant number / URL: 1SHD924N

ID: 206633

Start date: 01-10-2023

End date: 30-09-2027

Project abstract:

Colorectal cancer (CRC) is a major cause of cancer deaths. It develops from precancerous lesions, called polyps, which can be endoscopically detected and removed. Recent insights in the histomolecular features of the different types of polyps (hyperplastic polyp - HP, Adenoma - AD and Sessile serrated polyp - SSP) indicate that the classification of polyps and their associated risks differs significantly from the current classical adenoma-carcinoma sequence. Whereas ADs originate from WNT-dependent stem cell expansion, SSPs originate from a different neoplastic pathway, namely metaplasia, in which differentiated cells transdifferentiate into other, non-native celltypes. In SSP a pyloric-type gastric metaplasia has been observed, in which alien MUC5A-positive goblet cells are formed and colonize the colonic crypts from the colonic lumen down to the cryptbase. This fundamentally differs from ADs where stemcells from the crypt base expand towards the colonic lumen. Since metaplasia-based carcinogenesis is a top-down process, we hypothesize that artificial intelligence has the potential to help us discover the correlation between endoscopic image features and the underlying histomolecular features. This may empower a more accurate characterization and risk stratification in real-time during endoscopy and will help further unravel the pathways involved in colorectal carcinogenesis, leading to new strategies in prevention, surveillance, and therapeutics.

Last modified: 23-04-2024

Artificial intelligence for deep profiling and characterization of colorectal polyps
Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

- Generate new data:
  - Digital scans of histological slides. Data type .SVS. Expected total volume of data <2TB.
- Reuse existing data and generate new data:
  - Digital images in the form of still endoscopic images of colonic lesions (.PNG, .JPG, .TIFF) and endoscopic video's of lesions (.AVI, .MP4). Expected total volume of data >10TB
  - Demographic, clinical and histology data, collected from the medical history file of the patient (i.e. pathology diagnosis, LST grading, KUDO pitt pattern, age, gender, BMI, smoking, ...). Data type .CSV, .XLSX. Expected total volume of data <1GB

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

1. Designation of responsible person: Raf Bisschops
2. Storage capacity/repository: during and after the research, the data will be stored on the centrally managed UZ Leuven servers and KU Leuven onedrive with automatic back-up procedures and version tracking. There is no relevant limit as to the maximum size of stored data.

What’s the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

Not applicable.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

All existing and generated data comes from patients. Patients will be coded (i.e. pseudonymized), however there continues to be a link between the data and the individual who provided it. The subjects' identifiers will be stored separately from their researchdata and replaced with a unique code to create a new identity for the subject. This code is stored on the UZ Leuven server which is password protected, but which also allows to consult the electronic medical chart of the patient stored on UZ Leuven Hospital servers, only if necessary.

In addition, we will store all data on the central servers of the KU and UZ Leuven, which are protected against unauthorized access by firewalls.

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

Not applicable

Artificial intelligence for deep profiling and characterization of colorectal polyps
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	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
Endoscopic imaging (	Database	Generate new data and reuse existing data	Digital	*Observational*	*.TIFF* *.JPG* *.MP4* *.AVI*	*<50TB*	NA
Metadata	Database	Generate new data and reuse existing data	Digital	Compiled/aggregated data	.CSV .XLSX	<1GB	NA
Histology images	Database	Generate new data	Digital	Observational	.SVS	<5TB	NA

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:

During this research we reuse the endoscopic images and metadata which was collected in the CADartipod study.

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 clinical data, samples, and studies included in this project are approved by the Ethical Review Committee of the University Hospitals Leuven (S65253, S65254, S59405 & S64243)

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

Demographic and clinical data will be collected from the medical history file of the patients. All data will be pseudonymized.

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.

- Yes

Commercial valorization is possible trough the creation of computer-aided characterization tools. But this is not the primary aim of this project.

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.

- No

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).

All metadata will be collected in a central (per project) database and will follow a standard format and English vocabulary. All data will be pseudo-anonymized. Where deemed appropriate, a readme.txt file will be added to the research folder explaining data structure.

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.

- Yes

Yes, all stored metadata will follow a standard format and English vocabulary.

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

Where will the data be stored?

All data will be stored at either the KU Leuven Onedrive or the UZ Leuven servers (for clinical data). No data will be stored on local computers, hard drives etc.

How will the data be backed up?

All data stored in the central UZ/KU Leuven facilities are backed up automatically with version control and logging.

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

Data is stored either in the UZ Leuven data server. Currently this data server has ~ 43 TB storage space and can be further expanded as

necessary. Additionally the KULeuven onedrive has 5 TB total space.

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

The data will be stored on the UZ Leuven servers and KULeuven onedrive, which are only accessible with multiple factor authentication.

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

All costs are covered by the departmental group, or otherwise through existing grant 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...).

All data will be stored on the centrally managed UZ Leuven servers and KU Leuven Onedrive.There is no relevant limit as to the maximum size of stored data. Data will be stored for at least 25 years as requested by GDPR and UZ Leuven policies.

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

As mentioned above, all data will be stored on a centrally managed UZ Leuven data server and KU Leuven Onedrive.

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

All costs are covered by the departmental group, or otherwise through existing grant funding.

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:

Datasets will only be made available with third parties after a DTA.

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

Datasets will only be made available to third parties with a DTA, after prior aproval by prof Raf Bisschops (PI).

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.

- Yes, Privacy aspects
- Yes, Intellectual Property Rights

- Yes, Ethical aspects

All data originate from patients. Privacy regulations and ethical aspects restrict the sharing of these sensitive data.

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

A specific repository will be chosen after the publication strategy is known as some journal request specific repositories.

When will the data be made available?

If applicable for the research project, data will be made available upon publication of research results.

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

Data usage will either be open for public without any license in place, or be restricted and need a dedicated DTA before data sharing.

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?

Usually, there is no cost for sharing data with non-commercial third parties. For DTA of privacy-sensitive data, a quid pro quo in the form of co-authorship is usually requested.

6. Responsibilities

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

The endoscopy research team (Alexander Jans and study nurses: Hilde Willekens, Chelsea Camps & Anja Luypaert), under the supervision of Raf Bisschops

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

The endoscopy research team (Alexander Jans and study nurses: Hilde Willekens, Chelsea Camps & Anja Luypaert), under the supervision of Raf Bisschops

Who will manage data preservation and sharing?

The endoscopy research team (Alexander Jans and study nurses: Hilde Willekens, Chelsea Camps & Anja Luypaert), under the supervision of Raf Bisschops

Who will update and implement this DMP?

Alexander Jans & Raf Bisschops