DMP title

Project Name IOTA C3 DMP - DMP title

Grant Title C3/21/050

Principal Investigator / Researcher Wouter Froyman

Project Data Contact 016/342612 wouter.froyman@uzleuven.be

Description Ovarian cancer is the seventh most common cancer in women and constitutes the most lethal gynecological malignancy. Timely diagnosis and appropriate referral to gynecological oncology expert centers is pivotal to improve patient outcomes. Ultrasonography (US) is a readily available, cheap and harmless technique, and is widely accepted as first-line imaging modality for assessment of ovarian masses. Currently, the ADNEX model is the best available ultrasound-based mathematical model to differentiate between benign and several types of malignant ovarian tumors. However, it relies on the ability of the US operator to reliably (manually) locate, delineate and measure the tumor area and its associated features. Previously, we developed automated feature detection methods in collaboration with ESAT-STADIUS and General Electric, where we demonstrated the effectiveness of a deep convolutional neural network (DCNN) approach. The objective of this C3 project is to further develop a fully automated classification model for triaging ovarian cancer patients in non-specialty centers and to clinically validate them. Therefore, we perform an international multicentre prospective observational study of consecutive patients who are diagnosed with an adnexal mass. For this study, sites will store all relevant 2D ultrasound images and clips necessary to identify all the ultrasound features used in diagnostic models, and at least one 3D volume of the lesion for each subject enrolled.

Institution KU Leuven

1. General Information

Name of the project lead (PI)

Wouter Froyman, MD PhD

Internal Funds Project number & title

C3/21/050

IOTA-AI: AI-assisted automated detection of ovarian cancer on ultrasound imaging

2. Data description

2.1. Will you generate/collect new data and/or make use of existing data?

- Generate new data

2.2. What data will you collect, generate or reuse? Describe the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a numbered list or table and per objective of the project.

Type of data	Kind of data	File Format	Volume	Total amount	How created
Observational	Numeric and textual data; Clinical and ultrasound information	csv	Max 100 Mb	1 file	Data export from Clinical Data Miner (CDM), the study eCRF
Observational	2D ultrasound images	raw (proprietary format)	Max 8.8 GB (1.5-3.5MB per raw)	2500 2D Images	Pelvic ultrasound
Observational, converted	Converted 2D ultrasound images	tiff	Max 0.88 GB (0.15-0.35MB per tiff)	2500 2D images	via ExportRaw executable (provided by GE)
Observational	3D ultrasound volumes	vol (proprietary format)	Max 90 GB (8-90MB per volume)	1000 3D Volumes	Pelvic ultrasound
Observational, converted	Converted 3D ultrasound volumes	dcm	Max 90 GB (8-90MB per volume)	1000 3D volumes	via 4DView software (provided by GE)

3. Ethical and legal issues

3.1. Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.

No personal data will be used.

All csv-data is pseudonymized and can not to be associated to a person. CDM creates an additional anonymization layer (CDM-case- id) which makes it impossible to retrace the data to the actual person. No personal data is present in the csv.

All the data from the ultrasound images is completely anonymized. No personal data is mentioned on these images or metadata if the images. The images have the raw format which is stripped of any personal data.

3.2. Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s).

The project is an observational study, as such there is no deviation from the normal clinical routine.

We received approval from the EC UZ/KU Leuven, reference number S592047/B322201629748/AMEND-Id: 0003

3.3. Does your research possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

At the moment, there are no AI based algorithms clinically available for classification of ovarian cancer in patients based on ultrasound data but nonetheless there is increasing interest in the research community. Deep learning approach was first reported by Wu et al (*Advances in Multimedia Information Processing* - PCM 2018) and recently, a paper was published by a Swedish group (Christiansen et al, Ultrasound Obstet Gynecol 2021), that use deep learning to discern benign and malignant ovarian tumours. However, our proposed solution will use high-quality multi-centre real world clinical US data, which is unique in its completeness and novelty. Additionally, we will use raw 2D and 3D US images instead of processed 2D JPEG images used in the other studies, which will likely allow us to build more refined features.

Part of the previous work of the IOTA-AI project was done in collaboration with GE. However, the contract does not give GE any exclusive rights to our data or the developed models.

In addition, we performed a **patent search,** for a preliminary **freedom to operate** (FTO) analysis. The search was done in the European Patent database (EPO) for European granted patents. The chosen keywords were relevant for an AI/ML based algorithm based on ultrasound images, with use in ovarian cancer. 82 patents were found. After filtering out inactive patents, those regarding a specific chemical biomarker, compound, devices or apparatus for testing, we did not have any patents left. Therefore, our preliminary analysis of the patent literature indicates that we have a **full FTO and** anticipate no blocking issues for the foreseen valorization trajectories. However, we acknowledge that a full FTO may be necessary for confirming our position at a later stage. As APIs and software products are very dynamic in nature, we do not foresee a big opportunity to pursue for patents.

3.4. Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions regarding reuse and sharing are in place?

There are no restrictions for dissemination or exploitation of the data that we use in our project.

4. Documentation and metadata

4.1. What documentation will be provided to enable understanding and reuse of the data collected/generated in this project?

Metadata & appropriate documentation capturing information about the collected and generated datasets, features, and models will be made available.

Datasets:

1. The clinical and ultrasound metadata information can be extracted via CDM (study features described by the eCRF) where study features are (unambiguously) explained with a textual definition and an optional pictogram.

2. Ultrasound 2D images: The following information will be noted for each converted image (available in a separate .txt file per image): height, width, measurement specifications (boundary, origin, scale), scanner type & probe type.

3. Ultrasound 3D volumes: As the 3D volumes will be exported in DICOM format, the associated DICOM tags will contain the relevant meta-data about the 3D volume. These tags include but are not limited to: SOP Class UID, SOP Instance UID, Study date, #slices, spacing between slices, height, width ...

Features:

The extraction of the traditional radiomics features will largely adhere to the IBSI (Image Biomarker Standardisation Initiative) definition of features. For the higher-order deep learning features, documentation will be provided to allow for reproducibility.

Models:

Version control of the machine learning model will be provided by using GitLab KU Leuven (private repository). An accompanying ReadMe file will provide information about the environment configuration, the structure of the repository along with documentation about the different Python scripts (eg. hyperparameter configuration script, model script, optimizer script, train/test script, logger script...), the used resources for training (GPU, CPU) and visualization performances (according to specific hyperparameter setting).

4.2. Will a metadata standard be used? If so, describe in detail which standard will be used. If not, state in detail which metadata will be created to make the data easy/easier to find and reuse.

To handle ultrasound 3D volumes, the DICOM standard will be used. The 3D volume dataset is conveyed via the Enhanced US Volume SOP Class of the DICOM standard.

5. Data storage and backup during the project

5.1. Where will the data be stored?

1. In Europe. The csv file, images and volumes will be kept on our research unit (ESAT) central storage. Copies can be made on a personal device but it is not the intention to keep it there. When the data is not needed anymore on the hard drive, it should be removed.

5.2. How will the data be backed up?

The data will be stored on ESAT central servers with automatic daily back-up procedures.

5.3. 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, the storage & backup capacity of the esats centrals servers are sufficient for this dataset.

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

No additional costs.

5.5. Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Data security and access is handled by the DPO @ ESAT. When the data is stored, the DPO will define who has access to read and/or execute the data. No write access is permitted by those persons.

Only the people that have been granted access to the data will be able to access the data.

6. Data preservation after the end of the project

6.1. Which data will be retained for the expected 10 year period after the end of the project? If only a selection of the data can/will be preserved, clearly state why this is the case (legal or contractual restrictions, physical preservation issues, ...).

All relevant data will be stored for a mimum of 10 years.

- clinical and ultrasound information
- 2D images and 3D volumes
- all codes necessary for data analysis

6.2. Where will these data be archived (= stored for the long term)?

The data will be stored on the university's central servers (with automatic back-up procedures) for at least 10 years, conform the KU Leuven RDM policy.

6.3. What are the expected costs for data preservation during these 10 years? How will the costs be covered?

The database of 2D and 3D images, that will be compiled to realise the study will be hosted on the servers of ESAT- KU Leuven. There are no expected costs to do this. If there would be costs, these will be covered with internal funding.

7. Data sharing and re-use

7.1. Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions or because of IP potential)?

* Data is not the intended to be shared with 3rd parties.

The data-IP stays at KU Leuven.

(cfr protocol)

7.2. Which data will be made available after the end of the project?

The data will not be made available (cfr 7.1)

7.3. Where/how will the data be made available for reuse?

- Other (specify):

not shared

7.4. When will the data be made available?

not shared

7.5. Who will be able to access the data and under what conditions?

1, other researchers

* in own center

* in collaboration with other parties: with data transfer agreements (under condition of having an agreement with KU Leuven)

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

0

8. Responsibilities

8.1. Who will be responsible for the data documentation & metadata?

Wouter Froyman (wouter.froyman@kuleuven.be), the promotor of the project

8.2. Who will be responsible for data storage & back up during the project?

Willem Mestdagh (willem.mestdagh@kuleuven.be), software engineer from ESAT-STADIUS KU Leuven

Axel geysels (axel.geysels@kuleuven.be), PHD working on the data.

8.3. Who will be responsible for ensuring data preservation and sharing?

Wouter Froyman (wouter.froyman@kuleuven.be), the promotor of the project

8.4. Who bears the end responsibility for updating & implementing this DMP?

The end responsibility for updating and implementing the DMP is with the supervisor (promotor), Wouter Froyman (wouter.froyman@kuleuven.be)