"Reproducibility with VIP" project DMP

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - DMP template (english)" fourni par Agence nationale de la recherche (ANR).

Plan Details

Plan title	"Reproducibility with VIP" project DMP
Version	First version
Fields of science and technology (from OECD classification)	Health sciences, Computer and information sciences
Language	eng
Creation date	2022-06-08
Last modification date	2022-06-08
Identifier	ANR-21-CE45-0024
Identifier type	local identifier
License	Creative Commons Attribution 4.0 International

Project Details

Project title	Reproducibility with VIP
Acronym	ReproVIP

Abstract

In the last few years, there has been a growing awareness of reproducibility concerns in many areas of science. In a recent study, the analysis of a single neuroimaging dataset by 70 independent analysis teams reveals substantial variability in reported results, with high levels of disagreement across teams of their outcomes on a majority of tested pre-defined hypotheses. Despite the increase in awareness and a growing number of projects tackling this lack of reproducibility and proposing various tools to improve it, researchers still lack an integrated, end to end solution, providing a good level of reproducibility at a reasonable effort. In this context, ReproVIP aims at evaluating and improving the reproducibility of scientific results obtained with the Virtual Imaging Platform (VIP) in the field of medical imaging. We will focus on a reproducibility level ensuring that the code produces the same result when executed with the same set of inputs and that an investigator is able to reobtain the published results. We will investigate reproducibility at three levels: (L1) the code itself, and in particular different versions of the same code, (L2) the execution environment, such as the operating system and code dependencies, parallel executions and the use of distributed infrastructures and (L3) the exploration process, from the beginning of the study and until the final published results. At Creatis, since 2011, we have developed and deployed VIP, a web portal for medical simulation and image data analysis. By effectively leveraging the computing and storage resources of the EGI federation,VIP offers its users high-level services enabling them to easily execute medical imaging applications on a large scale computing infrastructure. In 2021, VIP counts more than 1200 registered users and about 20applications. In the last few years, VIP has addressed interoperability and reproducibility concerns, in the larger scope of a FAIR (Findable, Accessible, Interoperable, Reusable) approach to scientific data analysis. By implementing the CARMIN API and by using the Boutiques cross-platform framework for applications, VIP provides interoperability with existing platforms, which contributes to reproducibility. VIP provides us with a strong experience and a solid set of users and applications based on which we will tackle the lack of reproducibility L1, L2 and L3 described above. In order to reconstruct and interpret medical images, researchers make use of numerous image processing algorithms. Each processing step, from the raw image to the final decision, has its specific parameters and may come from a large number of different software packages and dependencies. As a result, the barrier to entry for non-expert users is high and can easily lead to processing pipelines quickly put together that are non-reproducible. Our final aim is to provide an integrated, end to end solution, allowing researchers to launch reproducible executions in a transparent manner. The proposed solutions for evaluating and improving reproducibility will be integrated in VIP and demonstrated on two scientific use-cases sharing a common set of processing tools for MRI image processing and addressing two different challenges: (i) optimising the MRI acquisition protocol w.r.t. to the signal to noise ratio (SNR) and (ii)optimising a processing pipeline for stroke prediction.

Funding

- French National Research Agency : ANR-21-CE45-0024

Start date	2022-02-01
End date	2024-01-31
Partners	• Institut Pluridisciplinaire Hubert Curien - IPHC (UMR 7178) (200612557C) • Concordia University / Big Data Infrastructures for Neuroinformatics ()

Produits de recherche :

1. Default research output (Dataset)

Contributeurs

Nom	Affiliation	Rôles
POP Sorina	CENTRE DE RECHERCHE EN ACQUISITION ET TRAITEMENT DIMAGES POUR LA SANTE	• DMP manager • Personne contact pour les données • Project coordinator

"Reproducibility with VIP" project DMP

1. Data description and collection or re-use of existing data

1a. How will new data be collected or produced and/or how will existing data be re-used?

Data used in task 3.1: Optimizing the MRI acquisition protocol

In the first part of the project, we will re-use preclinical data previously acquired in a study financed by Neurodis. Data concerns 30 rats followed in time during 6 months.

In the second part of the project, we plan to acquire new preclinical data on healthy subjects.

All the data has been and will be acquired on the PILoT imaging platform at Creatis, then stored in a Girder warehouse at Creatis.

Data used in task 3.2: Optimising a processing pipeline for skull base tumor segmentation

The use case will use existing data (that of the BRATS challenge in the form of a pre-trained model) to then test its transfer to new pre-processed data according to different versions of a processing pipeline. These new data will be clinical data from medical partners, stored on the Girder warehouse at Creatis after anonymization and not shared outside the project.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Data used in task 3.1: Optimizing the MRI acquisition protocol

For this use-case, data come mainly in two formats: (i) DICOM standard (including raw data and metadata) and (ii) a Bruker (constructor) proprietary format, containing a text (metadata) and a binary file.

Data acquired for une subject may include multiple acquisitions (at different times during 6 months) and may go up to a size of 1 GB. The data collected correspond to the following sequence acquisition : localized MR spectroscopy at short echo time (STEAM, TE 3ms) and DTI acquisition (EPI, 30 directions) in the rat brain at 11.7T

The processing of the data will involve the following format transformation: Bruker format to mrui file format for the MR spectrosocpy data, Dicom to MRTRix3 file format (.mih)

Data used in task 3.2: Optimising a processing pipeline for skull base tumor segmentation

For this use-case, data will be in two formats : (i) DICOM standard and (ii) NifTi format specifically for the representation and processing of n-dimensional data. The clinical data set will include 80 patients, each of whom underwent 2 MRI examinations (T2 and T1 with Gadolinium injection). A Gadolinium T1 MRI scan corresponds to about 170 MB while a T2 MRI scan to about 150 MB. The analysis of these data will produce new data in the form of binary masks corresponding to the semantic segmentation of tumors of the base of the skull.

---

2. Documentation and data quality

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Data used in task 3.1: Optimizing the MRI acquisition protocol

Metadata are extracted from the DICOM headers, but are also collected at acquisition time based on information provided by the operator.

These pieces of information, linked to imaging data, such as weight, age, duration of scan under anesthesia, will contribute to data production
for reproducibility and reuse and participation in population imaging or replication studies in preclinical domain.

Data used in task 3.2: Optimising a processing pipeline for skull base tumor segmentation

No metadata associated with the data will be stored or shared.

2b. What data quality control measures will be used?

Data used in task 3.1: Optimizing the MRI acquisition protocol

Data quality will be assessed at processing time, using the usual metrics and assessments such as SNR, linewidth for spectra, SNR, image distorsion for DTI.

Data used in task 3.2: Optimising a processing pipeline for stroke prediction.

Data quality will be assessed at processing time. Particular attention will be paid to ground truth and the contouring of neuronal lesions (labels or annotations) to be found via machine learning

---

3. Storage and backup during the research process

3a. How will data and metadata be stored and backed up during the research?

CREATIS will provide and manage a database infrastructure specifically developed to conduct studies on medical imaging. The warehouse, based on the Girder technology, will allow unique, private, secure and selective access. This access to the warehouse will be available through the web and through a REST API. CREATIS expertise is based on previous developments that use the warehouse technology for different national and international cohorts and actions.

The back-up system will ensure the availability of data during the time of the project, with a backup on a regular basis.

Concerning data from task3.2, a declaration to the DPO hase been performed with respect to the RGPD.

3b. How will data security and protection of sensitive data be taken care during the research

The Girder warehouse provides secure access through a web interface and a RESTful API. Access rights are configured according to the data protection requirements of each data collection. We can thus have stricter access rules for the second use case.

Data used in task 3.1: Optimizing the MRI acquisition protocol

For this use-case, these is no sensitive data (subjects are rats).

Data used in task 3.2: Optimising a processing pipeline for skull base tumor segmentation

Data series will be anonymized at our partners' institutions according to standard guidelines.

---

4. Legal and ethical requirements, code of conduct

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

Data used in task 3.1: Optimizing the MRI acquisition protocol

For this use-case, these is no personal data (subjects are rats).

Data used in task 3.2: Optimising a processing pipeline for skull base tumor segmentation

The project will rely on data from real subjects, collected in agreement with the declaration of Helsinki under informed consent from the studied subjects and agreement of the Ethics committees from our collaborators' institutions, anonymized according to standard guidelines, and accessible during the course of the project under restricted access, after signing a "Data use and distribution agreement" issued by the manager of the datasets at our clinical collaborators' institutions.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Data used in task 3.1: Optimizing the MRI acquisition protocol

The CNRS own the data.

Data used in task 3.2: Optimising a processing pipeline for skull base tumor segmentation

The patient data belongs to the HCL. Moreover, the data is there to validate the approach but will not be "sold" as such. The same for the model.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

Data used in task 3.1: Optimizing the MRI acquisition protocol

Ethical issues and code of conduct w.r.t. acquisition protocols on animal are handled by the PILoT imaging platform.

The project referenced as APAFIS#25081-2020050712321671 v1 has been authorized the 11th of May 2020 by the French ministry of education, research and innovation to collect data from small animal in vivo imaging.

Data used in task 3.2: Optimising a processing pipeline for skull base tumor segmentation

Ethical issues and code of conduct w.r.t. human/patient acquisition protocols are supported by HCL Lyon and the associated ethics committees.

---

5. Data sharing and long-term preservation

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Data used in task 3.1: Optimizing the MRI acquisition protocol

Data will be shared through the Girder web portal when the first article using them will be published.

Data used in task 3.2: Optimising a processing pipeline for skull base tumor segmentation

Patient data will not be shared. Only the models, the data already available and the reproducibility test methodology will be.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

For both use-cases, data will be stored on the Creatis Girder warehouse.

We will keep all raw/acquired data for a minimum duration of 5 years. We will select published results for preservation along raw data.

5c. What methods or software tools are needed to access and use data?

For both use-cases, data will be stored on the Creatis Girder warehouse.

Data stored in the Girder warehouse is accessible through a web interface and a RESTful API.

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

For both use-cases, data will be stored on the Creatis Girder warehouse.

This is work in progress, but we envisage to associate a DOI (obtained on platforms such as Zenodo) to data collections stored on Girder.

---

6. Data management responsibilities and resources

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

Frédéric Cervenansky: Research Engineer at CREATIS and administrator of the Girder platform.

6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

The dedicated warehouse will be managed by Frédéric Cervenansky, member of the ReproVIP project. This activity has been planned in the project and we estimate it at approximately 16 work hours. The warehouse and backups are services provided by the Creatis laboratory for internal projects at no additionnal cost.