ReproVIP Report #1.1.1
Selected Metrics and Reproducibility Results

Gaël Vila[1], **Morgane des Ligneris**[1], **Axel Bonnet**[1], **Yohan Chatelain**[2], **Tristan Glatard**[2], **Carole Frindel**[1], **Hélène Ratiney**[1], **Sorina Pop**[1]

[1] Univ Lyon, INSA-Lyon, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, Lyon, France.
[2] Concordia University, Department of Computer Science and Software Engineering, Montreal, Quebec, Canada.

Abstract. *This paper reports the mid-term achievements of the ReproVIP project, regarding reproducibility measures that could be relevant to the Medical Imaging (MI) research area. Two publications have been produced in this framework and will be presented at the ISBI conference in April 2023.*

Résumé. *Ce document rapporte les avancées à mi-terme du projet ReproVIP sur les mesures de reproductibilité applicables en imagerie médicale. Ces travaux ont donné lieu à deux publications qui seront présentés en conférence internationale (ISBI) en Avril 2023.*

Contents

1. Introduction

1.1. Background: the Reproducibility Crisis and the ReproVIP project

As with all disciplines that today rely on scientific computing, medical imaging (MI) is facing a reproducibility crisis [1]. The increasing complexity of data processing methods weakens one's ability to produce the same results twice, by applying the same treatments to the same set of inputs. Beyond the trivial influence of the analysis workflow [2], there is mounting evidence that computing environments (*e.g.* library calls, OS kernels, hardware infrastructures) also play a significant role by adding numerical uncertainty [3].

Relying on distributed computing resources, the VIP platform[1] is concerned by such potential variability in its outcomes. Beyond these operational concerns, computational reproducibility is a major challenge for the dissemination of scientific models in the academic world according to the new standards of Open Science.

The ReproVIP project addresses this issue at every level of data analysis, from the exploration process to the computing environment. It is structured around two complementary goals: (i) evaluate the uncertainty of digital outcomes and (ii) enhance the reproducibility of scientific calculations through the VIP platform. This paper reports the project's mid-term advances on issue number (i).

1.2. Issue: Multiple Sources of Variability

Once a data set has been collected, a numerical result is the combined product of three elements that fully describe a computing task on the VIP platform.

1. An execution environment, *e.g.* the hardware, OS used to make the computation;
2. A pipeline, *i.e.* a given version of a scientific application;
3. A methodological choice applied to the pipeline, *e.g.* a parameter set.

Generally speaking, the **execution environment** cannot be controlled by the researcher. This is true inside and outside the VIP platform, *e.g.* when some analysis workflow (like a Jupyter Notebook) is shared between several teams. The environment is the main reason why a calculation cannot be assumed invariant over time and space, despite the technical solutions available to minimise this effect (*e.g.* containers).

As scientific software continues to evolve, the **pipeline version** is a known source of variability in numerical outputs. Yet, clinical applications need to ensure the conclusions drawn from these outputs remain reliable across pipeline updates. Likewise, preclinical applications are subject to software balkanization when separate research teams propose **competing algorithms** to solve the same problem. Ensuring these algorithms lead to equivalent conclusions is an important step towards clinical use.

A similar goal can be set regarding the methodological choices made to run the pipeline, such as the **processing steps** and the **parameter set**. MI applications, which often answer an ill-posed numerical problem, tend to be very flexible in this regard.

The impact of these sources of variability on a scientific calculation should be considered as part of the uncertainty associated with any numerical result. The ReproVIP project

[1] VIP (the Virtual Imaging Platform) is a web portal for the simulation and processing of massive data in medical imaging. The platform is free and open-source; it currently has more than 1400 users and offers about twenty applications to the academic world. It uses computing and storage resources from the EGI e-infrastructure to enable high-throughput computing.

aims at investigating these levels of uncertainty; and ultimately making this knowledge available to the academic world through a dedicated dashboard.

1.3. Purpose: Evaluating Computational Reproducibility on Two Use-Cases

Our task was to evaluate how far *similar calculations*, when repeated on the same dataset, can produce *similar results*.

- The "how far" involves defining valid types of reproducibility **metrics**,
- Repetition involves carrying out reproducibility **experiments**.

A reproducibility metric may simply measure the differences between repeated outputs from a given MI application. However, MI applications produce a wide variety of digital objects. These can be either 1D, 2D, 3D or 4D vectors, whose elements (*e.g.* parameters, voxels) can take either continuous, categorical or tensor values. To narrow down this diversity of materials, two research fields have been selected as use cases for the ReproVIP project:

(i) Metabolite quantification for **magnetic resonance spectroscopy** (MRS),
(ii) Image preprocessing for **brain tumour segmentation** (BraTS).

In magnetic resonance spectroscopy (**MRS**), "metabolite quantification" amounts to estimating a neurochemical profile (*e.g.* the concentration rate of certain neurotransmitters) in a region of interest. To date, this technique is mainly used in preclinical studies and requires standardisation in software and methodological choices [4, 5].

In brain tumour segmentation (**BraTS**), a widely used image preprocessing pipeline results from an annual competition of deep learning algorithms for delineating gliomas in brain MRI volumes: the BraTS challenge [6]. Such tumour segmentation algorithms are intended for clinical use, as an aid to diagnosis or surgery. However, the impact of the preprocessing steps on the decision making process still needs to be investigated.

2. Reproducibility Experiments

In both research areas, reproducibility experiments were carried out to control for the multiple sources of variability mentioned above.

2.1. Use-Case I: *In Vivo* Metabolite Quantification

2.1.1. The Pipeline

Metabolite quantification can be described as a curve-fitting process on the magnetic resonance signal, during which metabolite concentrations are expressed as fitting parameters [7]. The procedure is summarized in Figure 1 below.

A typical application of such algorithm is in the monitoring of a particular neurotransmitter concentration over a period of time. Neurotransmitters can only be monitored *in vivo*, which means that no ground truth can be used to check the accuracy of the estimated concentration rates. The technique is calibrated with ex-vivo samples (phantoms) of marcomolecules and metabolites, which serve as reference for in-vivo measurements.

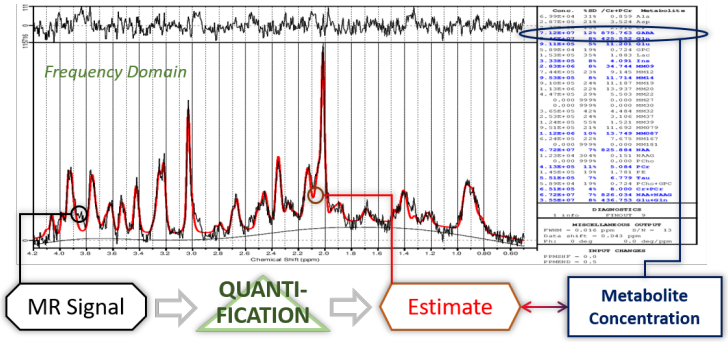

Figure 1. Metabolite Quantification in Magnetic Resonance Spectroscopy

2.1.2. Research issues

In metabolite quantification, reproducibility issues can been observed at three levels.

- High flexibility in the computation method (*e.g.*, model and parameters) leads to high variability in the final results [5].
- Beyond the chosen method, the analysis software leaves its own footprint [4].
- Within the same software, the non-deterministic parts of the fitting process lead to impactful variability in the numerical outcomes.

To our knowledge, however, no study to date has simultaneously controlled for these three sources of variability.

2.1.3. Experimental work

In the framework of the ReproVIP project, a study was made to capture these three sources of variability on the same MR spectroscopy dataset, by comparing the quantification outcomes: (i) between two quantification algorithms, (ii) between two sets of parameters for each software, and (iii) across multiple executions for each software-parameter set.

The study will be presented at the 2023 International Symposium on Biomedical Imaging. The paper can be downloaded below:

https://hal.science/hal-04006152

2.2. Use-Case II: Tumour Segmentation and the Preprocessing Pipeline

2.2.1. The pipeline

Before tumour segmentation, a pre-processing pipeline ensures that all brain scans are given a standardised shape [8]. This process goes through a number of intermediary states, as shown in Figure 2 below.

Finally, the DeepMedic inference model is used to delineate the tumour (right end of Figure 2). The outcome consists of a 3D masks, *i.e.* MRI volumes filled with 1s wherever the tumour is present and 0s elsewhere. Different parts of the tumour (*e.g.* necrotic core, oedema) are represented by as many masks, which are overlapped in the right end of Figure 2.

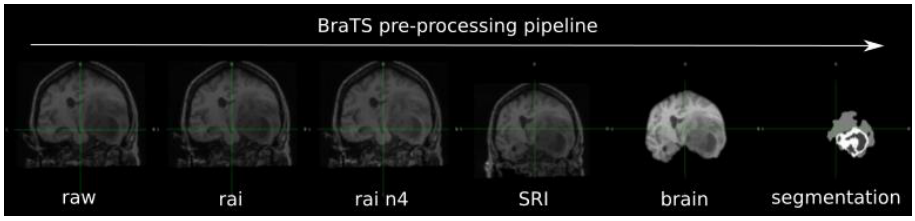

Figure 2. Tumour Segmentation and Preprocessing Pipeline

2.2.2. Research issues

Public pre-processing pipelines are now used as a standard routine. Anyone can access and use those pipelines on their own machines but little is known about the influence of the pipeline version or execution environment on the numerical results.

2.2.3. Experimental work

In the ReproVIP project, a second study was made to quantify the differences in results triggered by these two factors: (i) different versions of the same pipeline and (ii) numerical perturbations that mimic executions on different operating systems [9].

This study will also be presented at the 2023 International Symposium on Biomedical Imaging and can be downloaded below:

https://hal.science/hal-04006057

3. Selected Reproducibility Metrics : a Synthesis

3.1. Pipeline Outputs

Table 1 below summarizes the main output formats in both use cases. After metabolite quantification, the chemical profile (concentration rates) consist in a 1D vector of foating-point numbers. After the BraTS preprocessing pipeline, the structural MRI volumes consist in 3D volumes of foating-point numbers. After tumour segmentation, the results consist in separate 3D masks, *i.e.* 3D volumes with binary values.

Application (Use-Case)	Global Output	Local Value
Quantification (MRS)	1D Chemical Profile	Concentration $C \in \mathbb{R}^+$
Preprocessing (BraTS)	3D Brain Scan	Intensity $I \in \mathbb{R}^+$
Segmentation (BraTS)	3D Tumour Mask	Boolean $B \in \{0, 1\}$

Table 1. Numerical objects at the output of three MI applications.

3.2. Experimental Set

In both use cases, the selected metrics answered two types of reproducibility experiments. (i) When comparing two specific conditions (*e.g.* 2 pipelines, pipeline versions or parameter sets), one had to measure the *difference* between 2 similar results. (ii) When

comparing N similar conditions (*e.g.* N executions of the same pipeline with numerical perturbations), one had to measure the *dispersion* between N similar results. In both use-cases, metrics for dispersion and difference (or similarity) were chosen according to the proposed experiment: they are listed in Table 2.

Object (Use-Case)	Experimental Conditions	Selected Metric(s)
Concentration (MRS)	2 pipelines / 2 parameter sets	Test-Retest Statistics[2]
	N runs (random seeds)	Standard Error / C-R Bounds[3]
Brain Scan (BraTS)	2 pipeline versions	Peak Signal-to-Noise Ratio
	N runs (numerical perturbations)	Mean Nb. of Significant Digits
Tumour Mask (BraTS)	2 pipeline versions	Dice Coeff. / Hausdorff Dist.

Table 2. Metric choice as a function the object and experimental conditions.

3.3. Local or Global Scale

In Table 2, some metrics focus (a) on a single output value (*e.g.*, concentration rate), while others metrics include (b) a whole output object (*e.g.*, 3D tumour mask). The local scale (case a) is relevant when the researcher is interested in some parts of the pipeline's output (*e.g.*, some neurotransmitters). The global scale (case b) is useful when the whole result is of interest (*e.g.*, the whole tumour).

Metrics tailored for the local scale result in a reproducibility *profile*, with one result per component. This profile can be averaged to produce a global reproducibility value, such as the mean significant digit in the BraTS use-case.

3.4. Single Case or Multiple Input(s)

Finally, a reproducibility metric may focus either (a) on a single pipeline output, to evaluate a given *scientific result* ; or (b) on pipeline outputs from multiple inputs at the same time, to evaluate the *pipeline itself*. The single-case approach (case a) was implemented in the BraTS use-case, where inter-pipeline version and inter-pipeline run differences were quantified for each patient. The multiple-input approach (case b) was implemented in the MRS use case, where the test-retest metrics derived their statistical power from all signals included in the study.

Again, both approaches can be mixed: in the BraTS use-case, numerical instabilities were shown in the preprocessing pipeline by highlighting bad reproducibility results in the single-case approach. Conversely, metrics from a multiple-input experiment should be used to frame future reproducibility measures in the single-case approach. For example, by defining confidence intervals on the number of significant digits across multiple pipeline runs, or on the difference between two pipeline versions, for a given scientific result.

[2]Mean and confidence interval for the difference between both experimental conditions, across several paired samples.

[3]Cramèr-Rao bounds: theoretical bounds usually associated to least-square optimisation results

3.5. Conclusions

When setting a reproducibility experiment, the reference metric(s) should be selected by answering at least four questions:

(i) What kind of digital object is the pipeline output (*e.g.*, with continuous or categorical values);
(ii) Which analysis scale is relevant to interpret the pipeline output (*e.g.*, whole result or local value);
(iii) How many experimental conditions should be compared together (*e.g.*, two pipeline versions or numerous pipeline runs);
(iv) How many inputs will be used in this reproducibility experiment (*e.g.*, single result or full pipeline assessment).

Such variability in the experimental setting supports the use of distinct metrics to investigate pipeline reproducibility as a multi-faceted problem. Among the metrics proposed in Table 2, however, the number of significant digits can be highly generic:

- as a local metric, it can be averaged to make a global value;
- as a N-condition metric without need for statistical power, it can be used in the 2-condition problem;
- as a metric tailored for single results, it can be extended to a multiple-input experiment by generating confidence intervals.

In addition, its adimensional nature allows allows reproducibility measures to be compared between across different pipelines. For such reasons, the number of significant digits is a good candidate for establishing decision boundaries within automated reproducibility tests in the VIP platform. However, this metric is not well adapted to categorical objects like binary segmentation masks.

Finally, a fifth criterion could be added to the above list: the degree of familiarity of the pipeline's research community with the selected reproducibility metric. In the MRS use-case, academics were used to the Cramer-Rao bounds, a variance estimator; so the empirical variance was used in the reproducibility experiment for direct comparison. When designing a reproducibility dashboard for the VIP portal, frequently used metrics may be used as an aid for the user to understand reproducibility issues in their scientific results.

References

[1] Russell A. Poldrack, Chris I. Baker, Joke Durnez, et al., "Scanning the horizon: Towards transparent and reproducible neuroimaging research," *Nature reviews. Neuroscience*, vol. 18, no. 2, pp. 115–126, Feb. 2017.

[2] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, et al., "Variability in the analysis of a single neuroimaging dataset by many teams," *Nature*, vol. 582, no. 7810, pp. 84–88, June 2020.

[3] Tristan Glatard, Lindsay B. Lewis, Rafael Ferreira da Silva, et al., "Reproducibility of neuroimaging analyses across operating systems," *Frontiers in Neuroinformatics*, vol. 9, Apr. 2015.

[4] Alex A. Bhogal, Remmelt R. Schür, Lotte C. Houtepen, et al., "1H–MRS processing parameters affect metabolite quantification: The urgent need for uniform and transparent standardization," *NMR in Biomedicine*, vol. 30, no. 11, pp. e3804, 2017.

[5] Małgorzata Marjańska, Dinesh K. Deelchand, Roland Kreis, et al., "Results and interpretation of a fitting challenge for MR spectroscopy set up by the MRS study group of ISMRM," *Magnetic Resonance in Medicine*, vol. 87, no. 1, pp. 11–32, Jan. 2022.

[6] Spyridon Bakas, Mauricio Reyes, Andras Jakab, et al., "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," Apr. 2019.

[7] Josef Pfeuffer, Ivan Tkáč, Stephen W. Provencher, and Rolf Gruetter, "Toward an in Vivo Neurochemical Profile: Quantification of 18 Metabolites in Short-Echo-Time 1H NMR Spectra of the Rat Brain," *Journal of Magnetic Resonance*, vol. 141, no. 1, pp. 104–120, Nov. 1999.

[8] Spyridon Bakas, Chiharu Sako, Hamed Akbari, et al., "The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: Advanced MRI, clinical, genomics, & radiomics," *Scientific Data*, vol. 9, no. 1, pp. 453, July 2022.

[9] Ali Salari, Yohan Chatelain, Gregory Kiar, and Tristan Glatard, "Accurate simulation of operating system updates in neuroimaging using Monte-Carlo arithmetic," *arXiv:2108.03129 [eess, q-bio]*, Aug. 2021.