Fundamentals on Transparency, Reproducibility and Validation

Sorina Camarasu-Pop, Gaël Vila, Tristan Glatard, Axel Bonnet, Carole Frindel, Hélène Ratiney

Chapter 1

Fundamentals on Transparency, Reproducibility and Validation

Chapter Subtitle

Sorina Camarasu-Pop[a], Gaël Vila[a], Tristan Glatard[b], Axel Bonnet[a], Carole Frindel[a], and Hélène Ratiney[a]

[a]Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, FâĂŘ69100, Villeurbanne, France, [b]Department of Computer Science and Software Engineering, Concordia University, Montreal, Quebec, Canada

ABSTRACT

Transparency, reproducibility, and validation are fundamental concepts in research. Their definitions may vary among research disciplines (sometimes even lacking global agreement), but they all share common elements and practices. This chapter introduces the three concepts. Then, after discussing their definitions and interconnections, it illustrates their role in three main components of medical imaging studies — analyses, software, and data. For the three components, the chapter introduces methods and practical tools such as cross-validation, pre-registration, notebooks, code and data sharing, containerization, continuous integration, test-retest analysis, data quality, and challenges. Finally, some of these tools are illustrated through a concrete example from an ongoing initiative.

KEYWORDS

Transparency, Reproducibility, Validation, Software, Medical data analysis, Open science

1.1 INTRODUCTION

In scientific research, transparency, reproducibility, and validation are fundamental for ensuring the integrity and reliability of research results. As we explore these fundamental principles, it is essential to recognize the multifaceted nature of these terms. Each term unfolds along a different dimension, bringing a unique perspective to the overall integrity of the research effort. Although exhaustive definitions exist, the aim of this chapter is not to describe every nuance, but rather to highlight the general concepts and complexities inherent to their application.

Transparency advocates openness and clarity in research processes, inviting researchers to document the subtleties of their methodologies. Reproducibility calls for independent verification, underlining the need for others to reproduce

results to reinforce their credibility. Finally, validation examines the robustness and accuracy of results, ensuring that they conform to established standards.

These principles are especially important in medical image analysis. Research efforts in this field have identified unique challenges and considerations. The mix of technological advances and medical complexities requires a nuanced understanding of how transparency, reproducibility, and validation manifest in image analysis for medical purposes.

This chapter serves as a gateway to these essential concepts, highlighting the difficulties encountered, evolving definitions, and practical applications in the dynamic landscape of medical image analysis.

1.2 DEFINITIONS

1.2.1 Transparency

Transparency in research refers to the openness and clarity with which the research process and its outcomes are communicated and made accessible to others. Its main goal is to allow *a critical reader to evaluate the work and fully understand its strengths and limitations* [1]. This generally requires *the full disclosure of the research design, which includes the methods used to collect and analyze data, the public availability of both raw and manipulated data, in addition to the computational scripts employed along the way* [2].

In artificial intelligence (AI), transparency presents unique challenges. As [3] points out, simply publishing an algorithm's text or source code is not always sufficient for full understanding. This limitation becomes particularly evident for certain algorithms, notably those rooted in machine learning, where a holistic understanding is intimately tied to the datasets used for training. In addition to this limitation, the black-box nature of most AI systems renders transparency even more challenging. Explanations of machine learning and AI results have been proposed to mitigate such transparency issues [4].

The exact information which needs to be provided for thorough understanding and evaluation depends on each study, but at a general level we identify three key elements of transparency: share code, data and documentation.

Share code. Medical imaging research extensively uses numerical tools and methods for data acquisition, analysis, or sharing. Software code is thus the most accurate source of information for all the steps of a study. [5] considers that *the first and foremost strategy available to maximize the transparency of research methods is openly sharing the code with the minimal restrictions possible*.

Share data. Data is a central component of medical imaging research. Sharing both raw and derived data is essential for transparency. In the field of MI, however, ethical and privacy constraints may hinder data accessibility, so a detailed description of the data has to be provided.

Document choices and analyses (methodology). Clearly describing the research design, data collection methods, and analytical procedures enables others to understand how the study was conducted and assess the findings' validity and reliability.

Beyond these three elements that will be further discussed and illustrated in the following sections, transparency is also important with respect to:

- Conflict of interest and funding sources. Disclosing potential conflicts of interest (financial or personal) is important for understanding potential biases and influences on the research.
- Publication practices and peer review process. Information about the publication practices and peer review process contributes to the study's credibility.

Transparency enhances the credibility of research, fosters trust within the scientific community and allows for the effective evaluation and application of research findings. Transparency supports reproducibility and validation by providing all the necessary information about a study.

1.2.2 Reproducibility

Reproducibility can be seen as an umbrella term encompassing multiple terms, such as replicability, repeatability, reusability, and reproducibility itself, referring to the ability to recreate scientific results. [6] defines reproducibility as a spectrum of concerns that starts at a minimum standard of *same data* + *same methods* = *same results* to *new data and/or new methods in an independent study* = *same findings*. While there is no global agreement on the use of the terms *reproduce* and *replicate* for these two sides of the spectrum, this chapter adopts the definition in [7]:

- *Reproducible* research implies that authors provide all the necessary data and the computer codes to run the analysis again, re-creating the results.
- *Replication* is achieved when a study arrives at the same scientific findings as another study by collecting new data (possibly with different methods) and completing new analyses.

These definitions can be illustrated and extended in Figure 1.1 adapted from the Turing Way Community[1]. In this context, a study is reproducible when the *same analysis steps* performed on *the same dataset* consistently produce the *same results*.

Taking this definition further, one may inquire about the precise meaning of *the same* data, analysis, or results. For example, one could argue whether using random seeds or different libraries/software versions satisfies the *same* or *different* conditions. Results may also vary from bit-wise reproducible results to confidence intervals. This chapter considers that the answer to these questions

1. https://github.com/the-turing-way/the-turing-way

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalizable

FIGURE 1.1 Reproducibility matrix

depends on the study one tries to reproduce, which should mention what is to be considered the same data, analysis, or results.

Similarly, in [8], authors define *exact reproducibility* as the reproduction of strictly identical results as those of a previously published paper, i.e., being able to reproduce tables and figures as they appear in the original paper following the same procedures as the authors. In this case, "the same result" corresponds to what is reported in the paper (e.g., individual exact values, confidence intervals, or higher-level conclusions), which depends, in turn, on the appropriate validation method.

In addition to exact reproducibility, [8] also defines statistical reproducibility as the reproduction of the results of a study under statistically equivalent conditions, e.g., using another sample of data drawn from the same population. In this case, the results should be statistically compatible but do not need to be identical. Statistical reproducibility can also be part of the validation process described in the following section.

Reproducibility is considered a cornerstone of the scientific method, as it helps to establish the credibility of research and contributes to the cumulative nature of scientific knowledge. As seen so far, reproducibility can refer to a large spectrum of concerns. Depending on the study, their relevance may not be the same. For instance, exact computational reproducibility can be very important for detecting errors or differences and checking the robustness with respect to environments. However, a certain degree of variability may be necessary for certain studies, such as modeling physical uncertainties, making it impossible to achieve bitwise computational reproducibility. Therefore, addressing issues related to reproducibility requires a global view of the problem and efforts to improve research practices beyond exact reproducibility.

1.2.3 Validation

In the common language, validation is the process of making something officially or legally acceptable or approved, but also proof that something is correct. Validation in research refers to assessing and confirming the accuracy and reliability of research methods and findings. It involves demonstrating that the measures used in a study are appropriate and that the results are meaningful and trustworthy. As explained in [9], the word validation may have different meanings depending on the discipline. In software engineering, it means as-

sessing whether or not a particular system fulfills its intended purpose, which is contrasted with terms such as verification. In machine learning, conversely, validation means assessing how performant a system is on previously unseen data.

Despite these differences, validation procedures often share common elements, such as validation metrics. In [10], authors highlight the importance of metrics and their appropriate selection and propose a framework[2] for problem-aware metric recommendations. They explain that choosing the right metric is particularly challenging in image processing because the suitability of a metric depends on various factors, such as the (in)appropriate choice of the problem category (e.g., confusing object detection with semantic segmentation).

Validation metrics often involve comparison with a reference or ground truth. Acquiring an accurate reference or ground truth can be a major challenge due to variability among experts and potential errors made by them [11]. Subjective perception can lead to discrepancies between experts, emphasizing the need for strict protocols and clear definitions. Additionally, the complex nature of certain medical conditions can make it challenging to establish an unquestionable ground truth, as even experts may have different interpretations. These factors underscore the importance of carefully considering the nature of the reference or ground truth used in the validation process and acknowledging its potential limitations.

As mentioned in [10], beyond the "correctness" of an algorithm on a given set of test cases, there should be a holistic assessment including robustness and the ability to perform as well across different data sets (different protocols, different distributions, etc). [12] distinguishes between internal, temporal, and external validation for prediction models. Internal validation uses the patients from the development population, *i.e.*, the same data from which the model was derived. The most well-known forms of internal validation in machine learning are split-sample, cross-validation, and bootstrapping. Temporal validation uses data from the same study but is sampled at a different time interval than the data used for building the model. External validation includes patients that may differ from the development population in different ways (different countries, different types of care facilities, or different general characteristics). External validation provides thus evidence of the generalizability to various patient populations.

Validation is thus a multi-layer, never-ending process. The remainder of this chapter tackles aspects of evaluation rather than validation as a holistic assessment.

1.2.4 Global overview

As seen above, transparency, reproducibility, and validation encompass many principles for building accurate and precise models that meet the standards of

2. https://metrics-reloaded.dkfz.de/

Open Science. The concepts of accuracy and precision are typically depicted in Figure 1.2, where the optimum shot consistently lands in the center of the target (upper left diagram). They illustrate a possible distinction between validation and reproducibility.

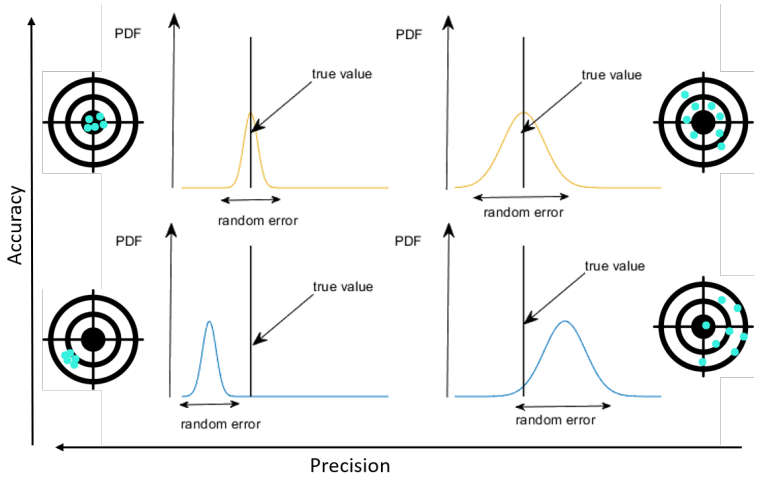

FIGURE 1.2 Accuracy and precision

From a broader perspective, each of the three concepts is a necessary but insufficient condition for producing a scientific investigation that others can replicate and improve. This is shown in Figure 1.3, which provides positive and negative examples of each concept (*e.g.* validated but non-reproducible or non-transparent work). This diagram also illustrates how deeply these three concepts are intertwined. For example, a successful replication can be seen as both a reproduction and a validation of the original study, the former being possible only if the latter is conducted with sufficient transparency.

These three concepts ultimately permeate every aspect of a scientific investigation. In computational analysis for medical imaging, these aspects can be captured in three main components:

- The analysis itself (*i.e.*, aspects that can be described in a scientific paper);
- The software tools (*i.e.*, scientific code and computing environments);
- The data.

The remainder of this chapter examines how transparency, reproducibility, and validation can be implemented in these three complementary elements of a typical 2020s research project.

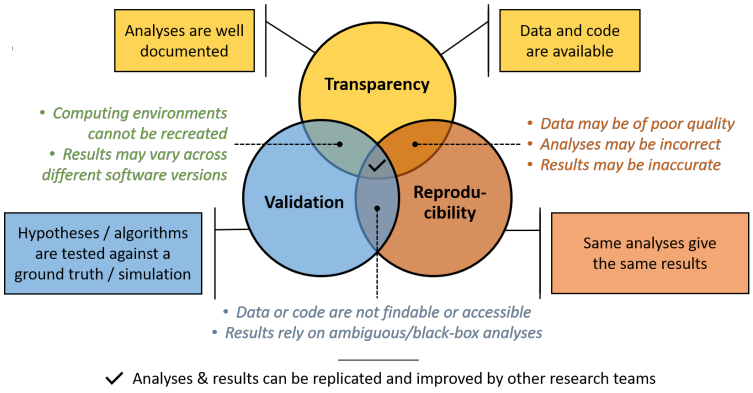

FIGURE 1.3 Transparency, reproducibility and validation as prerequisites for a replicable study

1.3 ANALYSES

The following section considers the means at the disposal of a researcher to evaluate and share a specific model or particular hypothesis. An *analysis* is anything that can be described in a scientific paper (or part of a paper), including the methods and algorithms used to produce such results.

1.3.1 Validation

Validation of a scientific analysis is a lifelong process that is not truly achieved until the study has been widely disseminated and replicated within the research community. The following paragraphs examine how a researcher can validate a model or hypothesis at their level by *evaluating* it against a dedicated *dataset*.

Evaluation strategies. In cases where real-world data are missing, simulation can be used as a substitute or complement (*e.g.* data augmentation). The evaluation strategy depends on the research goal using a properly curated dataset (see §1.5).

- Statistical tests evaluate a specific hypothesis regarding significance (p-value) and effect size. In this approach, particular attention must be paid to the sample size needed to analyze with sufficient statistical power [13].
- Validation metrics are used to assess the predictions of an existing model, focusing on performance and uncertainty. Such metrics should be carefully selected, as Section 1.2.3 explains.
- Cross-validation (CV) is used when the model is based on data-driven approaches such as machine learning.

Internal validity. Careful design of the CV strategy is crucial to the internal validity of a machine learning model. The process divides the data into three

distinct subsets: (i) the training set for model parameter learning, (ii) the validation set for hyperparameter tuning and model selection, and (iii) the test set for final model evaluation. The procedure is illustrated in Figure 1.4.

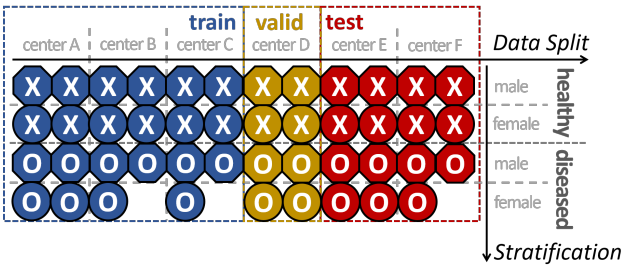

FIGURE 1.4 Dataset splitting for cross-validation: illustration on a multicentric medical dataset. The data is split group-wise across acquisition centers ("A"-"F") and stratified across classification labels ("healthy", "diseased") and patient sex ("male", "female").

To avoid data leakage effects [14], the test subset should be kept separate from the rest during the entire CV process[3]. Accurate model evaluation requires careful selection of the appropriate metric, depending on the intended application [10].

With this in mind, the training/validation steps can be iterated in various manners to enable model selection [15]. K-fold CV splits the dataset into k equally sized partitions, trains the model on $k-1$ partitions, and validates on the remaining one – the process is repeated k times. Monte-Carlo CV introduces a stochastic component by randomly sampling the training and validation sets over many iterations. The iteration strategy depends on the dataset's size and the time required for learning. According to [14], Monte-Carlo CV can provide confidence intervals on model performance, but statistical testing should never be used for model comparison. Finally, the splitting process should be planned carefully following two main principles (see figure 1.4): stratification and group-wise splitting.

Stratification means that each class (*e.g.*, disease state) should be equally represented in all subsets (training, validation, and test). This ensures that each subset is a representative sample of the original dataset. Ideally, the stratification step should also consider potential subclasses (*e.g.*, disease variant) to mitigate so-called "hidden stratification" effects [16] and account for relevant characteristics (*e.g.*, sex, age) that may cause unfair predictions from the model. Additional measures may be implemented when dealing with imbalanced datasets [17].

Group-wise splitting provides a more valid estimate of the model's generalization ability. In this approach, data points sharing common characteristics that should not be learned by the model (*e.g.*, same subject or same acquisition

3. This rule can be broken for small datasets using nested CV.

center) are grouped, and the groups are segregated when making the data subsets. This results in a better estimate of model performance on unseen data, as illustrated by [18].

External Validity. Group-wise splitting in CV assesses how a model can generalize in a given dataset. It cannot accurately estimate how the model will perform with the general public since a study *sample* cannot fully represent the target *population*. For instance, results for a mono-centric study may be subject to batch effects (see §1.5.3). This issue of external validity needs to be addressed before considering the output model for practical use.

In medical research, guidelines recommend testing the final model using data collected later or from different hospitals, countries, collection devices, protocols, and sociodemographic or clinical characteristics [14]. Blind assessment of generalization skills can also be achieved through research challenges (see §1.4.3).

As research datasets are often subject to quality control (see §1.5.3), field data from clinical routines should also be considered for robustness testing. Class prevalence should be considered when evaluating model performance, as it may vary between datasets and may not match the target population. This can be addressed with appropriate metric choice [10].

External validation is not specific to machine learning: testing a given model or hypothesis on different cohorts is a cornerstone of medical and human sciences. As this is a never-ending process, the most general guideline for increasing external validity is to make the study easily reproducible by others (see §1.3.3). Another fundamental way to externally validate a research work is through a peer-review process – this is a matter of transparency.

1.3.2 Transparency

Research Methodology. Looking back at the whole research process, the risk of false positive findings can be dramatically increased by the so-called "researcher degrees of freedom", which are well documented in the human sciences [19]. Researchers may conduct multiple tests until a significant result is observed ("*p*-hacking"), select data consistent with a particular hypothesis ("cherry picking"), or formulate hypotheses *post-hoc* based on the analysis outcomes ("HARKing"). Such behaviors, which undermine the integrity of the scientific process, can be driven by a variety of cognitive biases as well as unfortunate research incentives [20, 21]. In medical data analysis, they can be exacerbated by the flexibility of processing workflows [13, 22].

These challenges can be addressed by introducing proactive solutions into scientific practice – such as *pre-registration*, where key aspects of the study are registered before data are collected and analyzed [23, 20]. Online platforms

(such as the Open Science Framework and AsPredicted)[4] allow researchers to preregister their studies, promoting transparency and reducing the risk of *post-hoc* adjustments.

While originally designed for confirmatory studies (where hypotheses are made before data collection), this approach is being increasingly used and adapted for secondary data analysis [23]. This brings the practice within the reach of exploratory processes such as machine learning, which often rely on datasets not acquired for a specific purpose.

Registering a study can also be part of the publication process [24]. An increasing number of scientific journals (over 300 in late 2023, according to the Center for Open Science[5]) offer peer-review and paper acceptance based on the study design, *i.e.* before the results are known. Registered reports offer researchers a guarantee that the results will be published in the event of negative findings in return for a rigorous and transparent research plan. However, a given analysis cannot be fully reviewed or reproduced without transparent access to its practical implementation, even if properly documented in a classical or registered paper.

Implementation. Computational analyses rely on multiple, interdependent software layers. All of them are founded on lower-level implementations (*e.g.* `glibc`) and operated through a specific operating system on dedicated hardware, the sum of which can be called the computing environment. Efficient sharing of software components and computing environments is a delicate matter. This is discussed in Section 1.4.

At the analysis level, computational notebooks (*e.g.* Jupyter or R Markdown) allow the combination of documentation in natural language (*e.g.*, English) with snippets of macros and source code in a single, interactive report. They can be used to orchestrate an analysis from data curation to figure drawing, provided this can be scripted in a few languages. The exponential growth in their adoption (over 11 million Jupyter notebooks on GitHub in early 2023)[6] has made them a *de facto* new standard for scientific dissemination.

Guidelines for properly designing and sharing a notebook include code modularization and careful recording of all dependencies [25]. Comprehensive documentation of each analysis step is essential because of the increasing flexibility of modern data processing workflows. From this perspective, the transparency of an analysis is closely related to reproducibility.

4. `https://osf.io/registries` — `https://aspredicted.org/`
5. `https://www.cos.io/initiatives/registered-reports`
6. `https://github.com/parente/nbestimate`

1.3.3 Reproducibility

A typical example of reproducibility problems caused by analysis flexibility was presented in [22], where 70 research teams analyzed the same fMRI dataset to test the same hypotheses. Since no two teams designed the same analysis workflow, on average, 20% of the teams came to opposing conclusions on the nine tested hypotheses. Transparent sharing of analysis methods and scripts can help reduce such degrees of freedom (see §1.3.2), but rigorous reporting of a research methodology is far from trivial. This can be done according to institutional guidelines, such as a reproducibility checklist[7].

Another critical (but still largely unknown) issue relates to the computing environment. For example, [26] ran the same analyses (*e.g.* cortical thickness extraction with CIVET) on two separate computing clusters and found statistically significant differences between the two clusters. They could blame a variation in versions of `glibc` for these differences. To mitigate such effects and make "re-runnable" analyses, off-the-shelf solutions are available to couple a Jupyter notebook with a computing environment, such as Jupyter Binder or Google Colab[8]. These solutions rely on containerization (see §1.4), an efficient but still imperfect approach to exact reproducibility.

In neuroimaging research, open platforms such as Neurolibre[9] go the extra mile by making all the elements of an analysis available in one place. In such "reproducible preprints", the research paper can be accompanied by a dataset archive, a source code repository, a notebook and a Docker container.

Beyond analysis reporting, however, it should be noted that notebooks are not made for software development. In clinical applications, a full data processing workflow would involve a workflow manager such as NiPype [27], Snakemake, or Nextflow[10]. Scientific workflow systems automate the diverse but repetitive steps involved in *in-silico* experiments, from input/output data management to data analysis. They enhance reproducibility as shown in [28]. Workflows can also record provenance [29], preserving essential metadata such as where, when, and how data were produced.

The examples above show how any scientific analysis relies on the software and data used to carry it out. The following sections provide guidelines for using both components in a valid, transparent, and reproducible manner.

1.4 SOFTWARE TOOLS

Scientists spend 30% or more of their time developing software [30]. Medical image analyses rely on complex software ecosystems composed of (a) soft-

7. `https://miccai2021.org/files/downloads/MICCAI2021-Reproducibility-Checklist.pdf`
8. `https://mybinder.org/` — `https://colab.google/`
9. `https://neurolibre.org`
10. `https://www.nextflow.io/` — `https://snakemake.github.io/`

ware implementing study design and analysis, typically implemented as custom scripts or notebooks, (b) core image processing methods such as segmentation and registration, such as Freesurfer [31] or FSL [32], (c) direct software dependencies such as optimization toolboxes or data manipulation libraries, and (d) contingent dependencies such as elementary mathematical functions, compilers, interpreters, and other tools typically provided by operating systems. At all these levels, software tools implemented in various programming languages are commonly combined, including Python, C and C++, or Matlab. As explained in [33], software variations at these four levels can substantially impact imaging analyses. Therefore, transparency, validation, and reproducibility matter across the entire software ecosystem.

1.4.1 Transparency

Software tool transparency primarily implies source code availability. Over the past years, open-source development tools have matured and are widely adopted in research communities.

Most notably, Git[11] has emerged as a robust solution to manage source code and share it on online platforms such as GitLab or GitHub. Best practices for code sharing, including clear licensing, proper documentation, and code formatting standards, are available, and all support code transparency [34].

Software version control and releases are particularly important to accurately identify software tools. Platforms such as Zenodo[12] and the Open Science Framework[13] associate permanent identifiers (Digital Object Identifiers or DOIs) to software releases, which references them in the long term. Scientific software tools can also be assigned Research Resource Identifiers (RRIDs[14]) to further improve transparency. The work in [35] is an excellent example of transparent software reporting.

1.4.2 Reproducibility

The exact reproducibility of entire software ecosystems is often out of reach, given the breadth of software involved in image analyses. Therefore, some flexibility in the scope implied by *same code* is generally admitted, given the expected level of robustness of analyses to ancillary dependencies such as compilers or parallelization frameworks.

In this context, software reproducibility is commonly facilitated by publishing versioned software packages — for instance through the Python Package Index (PyPI), operating system repositories, or directly on GitHub — or by re-

11 https://git-scm.com
12 http://zenodo.org
13 http://osf.io
14 https://www.rrids.org

leasing software container images executable with the Docker[15] or Apptainer[16] engine. For instance, in neuroimaging, NeuroDebian [36] provides a collection of popular software packages for the Debian and Ubuntu operating systems, and NeuroDocker[17] facilitates the creation of Docker images containing common software tools.

Software packages and container images each have their own advantages and address a different trade-off between reproducibility and generalizability: software packages are generally more lightweight and transparent, whereas container images encompass a larger subset of the software ecosystem. Interestingly, the Guix [37] package manager provides extensive tracking of software dependencies — up to their compilation.

The current solutions still have important limitations regarding reproducibility. In particular, software tools are commonly compiled with hardware-specific options to leverage recent CPU architecture advances, making containers and packages less portable across execution environments.

1.4.3 Validation

Software testing. The validation of software tools is a field in itself, with different implications across software ecosystems. Software tests are widely used to assert software functionality, detect regressions across versions, and ensure portability across environments. Software tests are usually executed automatically throughout the software development process, using continuous integration (CI) tools such as GitHub actions, CircleCI, Jenkins, or other systems. Nibabel[18], ITK [19], and fmriprep[20] are excellent examples of medical imaging software projects using software tests with different programming languages.

The validity of software tools, however, involves considerations broader than can be captured using traditional software testing. In medical imaging, validity varies across datasets due to differences in subject populations and acquisition parameters. Besides, establishing evaluation references in in-vivo imaging is not straightforward due to the absence of ground truth. As a result, specific validation protocols have been defined, involving simulated data, human expert references, and a variety of imaging protocols and subject populations.

Challenges. Ultimately, validation requires an objective third party to mitigate the risk of overfitting software parameters to the validation dataset. For this reason, challenges are commonly organized in the medical imaging and ML communities, providing comparative evaluations of software tools tailored for

15.https://www.docker.com
16.https://apptainer.org
17.https://github.com/ReproNim/neurodocker
18.https://github.com/nipy/nibabel
19.https://github.com/InsightSoftwareConsortium/ITK
20.https://github.com/nipreps/fmriprep

specific tasks.

Illustrating this, the works in [38] and [39] delve into two software challenges organized by the MICCAI conference through the VIP platform. During these challenges, participants submitted their software tools encapsulated in Docker containers that the challenge organizers subsequently executed on an independent validation dataset.

These challenges assume a pivotal role in the large-scale validation of AI models, offering a dynamic and comprehensive testing ground across diverse applications. By presenting real-world scenarios and datasets that extend beyond conventional training sets, challenges facilitate the assessment of AI models' generalization capabilities. Within the structured frameworks of these challenges, standardized evaluation criteria create a common ground for researchers and developers to objectively benchmark different models. This, in turn, aids in the identification of effective approaches and driving advancements in the field, as evidenced by works like [40], [41].

Moreover, challenges contribute substantially to the practical validation of AI in real-world contexts, ensuring that models align with and perform well in various scenarios. Despite their undeniable benefits, challenges come with limitations, including variations in constraints on code availability and variations in result presentation formats. These nuances underscore the need for meticulous consideration in the validation process, as highlighted in [42].

1.5 DATA

Data undergo multiple changes from acquisition to final processing within a given study. The notions of transparency, reproducibility, and validation apply to each of the multiple steps within a data's lifetime. The following provides a non-exhaustive overview of some of the principles, guidelines, and tools addressing these concerns.

1.5.1 Transparency

As discussed in Section 1.2.1, transparency involves data sharing and documentation. However, In the field of medical imaging, ethical and privacy constraints may hinder the sharing and accessibility of data. Consent forms[21] should address these questions prior to data collection. When data access and sharing is not possible, proper description and documentation should be provided to ensure a certain degree of transparency. Metadata and documentation are essential to transparency, both when data can and cannot be shared.

Various guidelines, such as [43, 5, 44], are available in the literature to inform and help researchers follow best practices. The widely accepted FAIR

21 https://open-brain-consent.readthedocs.io/en/stable/

principles[22] also contribute to data transparency by guiding to make research data:

- Findable. Metadata and data should be easy to find, *e.g.* by using globally unique and persistent identifiers and indexing them in searchable resources.
- Accessible. Users need to be able to access the (meta)data, possibly including authentication and authorization.
- Interoperable. Data should use a formal, accessible, and shared language to facilitate the integration with other data, as well as with applications or workflows for analysis, storage, and processing.
- Reusable. Metadata and data should be well-described to be replicated and/or combined in different settings.

In this context, Data Management Plans (DMP) are key elements for good data management. They are formal documents outlining how data is handled during and after a research project. Many funding agencies require a DMP as part of their application processes. For example, the French National Research Agency (ANR) requires all projects funded since 2019 to produce a DMP. This can be done through the DMP OPIDoR[23] online service, which provides guidance through the drafting and implementation in practice of data or software management plans. DMPs address important questions concerning data description and collection or re-use of existing data, documentation and data quality, storage and backup, legal and ethical requirements, data sharing and long-term preservation, data management responsibilities, and resources.

Once these questions are addressed, multiple data management platforms can help store, share, and retrieve data- ideally in compliance with the FAIR principles. They can be dedicated to medical imaging, such as Shanoir [45], Loris [46] and Neurobagel, allowing for the implementation of specific features (e.g., support for ontologies or BIDS[24] data), or general-purpose repositories or warehouses, such as DataLad [47], Girder[25] and Zenodo.

1.5.2 Reproducibility

Regarding data, reproducibility concerns the ability to reproduce their generation and determine whether collecting/measuring them is not the result of chance but corresponds to a mastered and understood process.

A test-retest experiment can be carried out to qualify the reproducibility of a data acquisition process. Test-retest experiments consist of (i) repeating a measurement procedure (*e.g.*, a medical scan) on the same subject or sample within a short period and (ii) assessing the differences/variability between the repeated measurements. Then, an assessment is often performed using Bland-

22.https://www.go-fair.org/fair-principles/
23.https://dmp.opidor.fr
24.https://bids.neuroimaging.io/
25.https://github.com/girder/girder

Altman statistics or correlation. Thus, test-retest analysis can be found in all imaging modalities. Still, there is always room for discussion and improvement in this investigation, as it is wise to remember that promoting reliability should not be done at the expense of validity [48].

Such "metrological" considerations apply as soon as the acquisition system can be regarded as a measuring instrument. They can either relate to data acquisition alone or extend to post-acquisition processing, such as quantitative parameter estimation. These steps are essential to assess and find ways to improve the reliability of medical imaging data [49], and to make reliable clinical interpretations or research development.

1.5.3 Validation

Data quality. Validating the data used in a scientific study is first and foremost a question of "quality", *i.e.* determining whether the data are usable and correspond to the behavior expected in the analysis process. The goal here is again to warranty reliable and trustworthy data.

Data quality scores can be set up to check for artifacts due to motion blur or ghosting, which could affect the accuracy of the analysis. Rejection criteria need to be set up and clearly enunciated, knowing that objective criteria can be difficult to design and should sometimes rely on an expert's subjective perception [50]. Sometimes, data rejection, resulting in reduced data, can counter-intuitively help gain statistical power. Still, the reason for the rejection must be clearly defined and based on an objective data quality score [51]. Indeed, data selection significantly impacts what will be considered statistically "representative of a population" or "sufficiently general to capture diverse behavior patterns".

Multicentric datasets. A medical imaging study reaches a key *validation* stage when the methods implemented are successful on widely collected data – *i.e.* not just for one research site, but across multiple centers. Indeed, obtaining scientific results from a set of data from a single center or learning a model from data from a single type of machine can lead to misleading or biased results. This "batch effect" needs to be addressed or questioned. The challenge of validation on multi-centric data is to question the generalizability of a scientific method/result/conclusion, *i.e.* its capacity to be applied to a larger population or to conditions different from those initially studied (see also §1.3.1 about external validity).

However, the use of multi-center data is not without pitfalls. Data must be standardized to mitigate or eliminate the effects of different centers. Indeed, some variations in the data can come from differences in data collection procedures, equipment, demographics, and other factors across centers (actually, the same aforementioned batch effect) and should not be misleading. Along with multicentric data, the problem of their integration arises. The scientific community is actively exploring innovative methodologies such as the Combat method

and federated learning to address it. The Combat method, originally introduced in the field of genomic [52], is now used in medical imaging to "combat the batch effect" [53]. It is a harmonization technique to reconcile differences between data collected from different sources to ensure a coherent, unified data set. With a data-driven approach, it estimates the "site's" effect. Its application enables researchers to merge heterogeneous data while minimizing the risk of introducing biases that could compromise the generalizability of results. On the other hand, federated learning [54] is a cutting-edge approach in which models are trained collaboratively at decentralized sites without exchanging raw data. While this technique offers promising scalability and privacy preservation prospects, it also presents some challenges. Among these is the need for effective communication and coordination between distributed sites and the potential heterogeneity of local datasets.

In conclusion, a delicate balance must be found between (i) increasing the dataset size to improve the inferred model's generalization skills and (ii) overcoming the challenges inherent in harmonizing multi-centric datasets.

1.6 DISCUSSION

The following section concludes this chapter with a summary of the above-mentioned concepts and tools and a practical example of how some have been implemented in a recent project.

1.6.1 Summary

Table 1.1 sums up the main practices and tools presented in previous sections. The list is not exhaustive but gives a glimpse of the bigger picture and the wide range of existing tools. Given their diversity, it may be difficult for a young researcher to master them all quickly (*e.g.*, a PhD thesis) in addition to another main research subject. To facilitate the use of these diverse tools, initiatives such as Neurolibre offer high-level service solutions that combine different tools and facilitate their adoption.

1.6.2 Practical example

The ReproVIP[26] initiative aims at evaluating and improving the reproducibility of scientific results obtained on the Virtual Imaging Platform (VIP)[27]. This practical example attempts to bring together multiple solutions addressing transparency, reproducibility, and validation at different layers. It is not meant to reflect a perfect solution (since it is not) but rather one illustration of an existing initiative.

VIP is a free web portal for the analysis of medical imaging data. It renders scientific software tools accessible as a service by deploying them on distributed

26.https://anr.fr/Projet-ANR-21-CE45-0024
27.https://vip.creatis.insa-lyon.fr/home.html

TABLE 1.1 Summary table

	Transparency	**Reproducibility**	**Validation**
Analysis	Pre-registration, access to documentation and implementation, analysis notebooks (*e.g.* Jupyter, R)	Workflows (*e.g.* NiPype, Snakemake, Nextflow), open platforms (*e.g.* Neurolibre), reproducible preprints	Statistical tests (*e.g.* p-value), validation metrics, cross-validation, external validation
Software	Code versioning & sharing (Git), licensing & documentation, open platforms (*e.g.* Zenodo, OSF), identifiers (*e.g.* DOI, RRID)	Versioned software packages (*e.g.* PyPI), container images (*e.g.* Docker, Apptainer), Guix	Software testing, CI tools (*e.g.* GitHub actions, Circle CI, Jenkins), validation protocols, challenges
Data	Data & metadata sharing & documentation, FAIR principles, DMP, standards, data management tools	Test-retest, agreement analysis (*e.g.* BlandâĂŞAltman statistics)	Data quality, multi-centric data, statistical methods (*e.g.* Combat)

computing resources based on their Boutiques[28] descriptors. VIP offers both a web Graphical User Interface (GUI) and a REST API[29] allowing for interoperability and analysis automation.

Software tools. At the scientific software level, VIP handles their deployment and execution but is not involved in their development. VIP fosters transparency with the help of Boutiques[30], which, among others, allows to easily publish descriptors on open repositories, such as Zenodo, to make them findable and accessible (*e.g.* the BraTSpipeline descriptor[31]). Boutiques also leverages the use of containers associated with well-identified software versions.

Regarding reproducibility, since VIP uses distributed heterogeneous computing resources, it is particularly important to consider the execution environment. Almost all applications available in VIP are based on containers. The ReproVIP project also investigates using the Guix package manager to deploy applications on heterogeneous resources.

Since tools are developed externally, their validation is generally out of reach for VIP. Recently, however, a CI platform has been able to run tests on VIP automatically and regularly to verify that an application produces the expected results for known inputs. ReproVIP also introduces a web dashboard with visualization components to display and interpret different VIP results. For instance, it can be used to compare two images generated by two application versions over the same input, offering side-by-side visualization and appropriate metrics to estimate the differences in results.

28.https://github.com/boutiques/boutiques
29.https://github.com/CARMIN-org/CARMIN-API
30.https://figshare.com/articles/poster/fair-pipelines-poster_pdf/8143241
31.https://zenodo.org/records/7779113

Analyses. At the analysis level, transparency is fostered by providing Notebook templates[32] allowing the integration of the whole exploration process, including calls to the VIP API through the VIP client. If based on such Notebooks, analyses are rendered reproducible by their simple re-execution either locally or on platforms such as Binder. It should be noted that the execution of the software tools is handled by VIP through boutiques, thus ensuring a certain level of reproducibility. Since its latest version, VIP has allowed sharing a given experiment with the community. Subject to validation by VIP admins, this functionality allows us to push results and execution traces on Zenodo, as well as relaunch the experiment by another user (provided that he/she can access input data).

Data. VIP is a computing platform lacking many data-dedicated functionalities that data management platforms can provide. VIP handles data only temporarily for processing purposes and interconnects with data management platforms such as Shanoir and Girder for longer-term data management. At the CREATIS laboratory[33], multiple Girder warehouses are used. One of them is the PILoT warehouse[34] interconnected with the PILoT imaging facility[35] and VIP. Data acquired on PILoT can be automatically pushed to the warehouse with associated metadata (extracted from DICOM headers or additional information sources created at acquisition time). Both Girder and VIP provide RESTful APIs, facilitating their interconnection. Data stored on Girder can be thus processed on VIP and results stored back on Girder with processing metadata (see example available on the client github repository[36]). The whole process is illustrated on Figure 1.5.

As a reminder, results produced on VIP, along with processing metadata can also be exported to Zenodo for long-term storage and DOI retrieval for publication. Although not as detailed and powerful as a provenance system, it provides a customizable, easy-to-implement solution to enhance transparency and reproducibility. Flexible and customizable solutions can prove very useful for small independent projects, but standardization becomes essential when sharing at a larger scale.

1.6.3 Conclusion

This chapter discussed the essential concepts that should be found in all scientific work. They include the need for rigor and documentation, as well as

32 https://github.com/virtual-imaging-platform/VIP-python-client/tree/develop/examples
33 https://www.creatis.insa-lyon.fr/site/en
34 https://pilot-warehouse.creatis.insa-lyon.fr
35 https://www.creatis.insa-lyon.fr/site/fr/node/47253
36 https://github.com/virtual-imaging-platform/VIP-python-client/blob/develop/examples/bruker_preprocessing/preprocess.ipynb

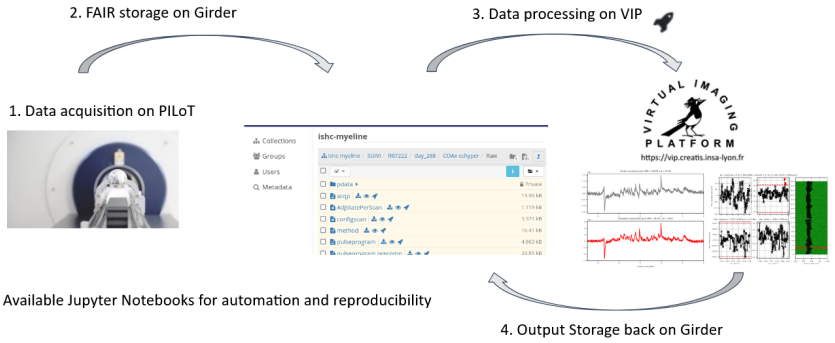

FIGURE 1.5 Illustration of an initiative of bridging data acquisition, storage, and processing solutions to enhance transparency and reproducibility.

the constant questioning of the reliability and trustworthiness of any generated results. Guidelines, consensus papers, recommendations, and requirements for reproducibility need to be drawn up at a time of massive data manipulation and staggering enterprise, as proposed by advances in artificial intelligence. However, giving fixed ways of doing things, definitions, and procedures should be handled carefully. Researchers should keep questioning and revisiting them, leaving space for thinking outside the box as sensitivity to new situations, variability, and the unknown. These key elements are part of the commitment to research.

1.6.4 Acknowledgements

This work was supported by the French ANR through the ReproVIP project (ANR-21-CE45-0024-01) and by the Canada Research Chairs program. This work was performed within the framework of the LABEX PRIMES (ANR-11-LABX-0063).

BIBLIOGRAPHY

[1] T. E. Nichols, S. Das, S. B. Eickhoff, A. C. Evans, T. Glatard, M. Hanke, N. Kriegeskorte, M. P. Milham, R. A. Poldrack, J.-B. Poline, B. Proal, Erikaand Thirion, B. T. T. Van Essenand White, Tonyaand Yeo, Best practices in data analysis and sharing in neuroimaging using mri., Nature neuroscience (2017) 299–303doi:10.1038/nn.4500.

[2] D. F. Filho, R. Lins, A. Domingos, N. Janz, L. Silva, Seven reasons why: A userâĂŹs

37.https://creativecommons. org/licenses/by/4.0/

guide to transparency and reproducibility, Brazilian Political Science Review 13. doi:10.1590/1981-38212019000200001.
URL https://brazilianpoliticalsciencereview.org/article/seven-reasons-why-a-users-guide-to-transparency-and-reproducibility/

[3] C. Henin, D. Le Métayer, Beyond explainability: justifiability and contestability of algorithmic decision systems, AI & SOCIETY (2021) 1–14.

[4] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, S. Kujala, Transparency and explainability of ai systems: From ethical guidelines to requirements, Information and Software Technology 159 (2023) 107197. doi:https://doi.org/10.1016/j.infsof.2023.107197.
URL https://www.sciencedirect.com/science/article/pii/S0950584923000514

[5] G. Niso, R. Botvinik-Nezer, S. Appelhoff, A. De La Vega, O. Esteban, J. A. Etzel, K. Finc, M. Ganz, R. Gau, Y. O. Halchenko, P. Herholz, A. Karakuzu, D. B. Keator, C. J. Markiewicz, C. Maumet, C. R. Pernet, F. Pestilli, N. Queder, T. Schmitt, W. SÃşjka, A. S. Wagner, K. J. Whitaker, J. W. Rieger, Open and reproducible neuroimaging: From study inception to publication, NeuroImage 263 (2022) 119623. doi:https://doi.org/10.1016/j.neuroimage.2022.119623.
URL https://www.sciencedirect.com/science/article/pii/S1053811922007388

[6] L. A. Barba, Terminologies for reproducible research (2018). arXiv:1802.03311.

[7] R. D. Peng, Reproducible research and Biostatistics, Biostatistics 10 (3) (2009) 405–408. arXiv:https://academic.oup.com/biostatistics/article-pdf/10/3/405/26055640/kxp014.pdf, doi:10.1093/biostatistics/kxp014.
URL https://doi.org/10.1093/biostatistics/kxp014

[8] O. Colliot, E. Thibeau-Sutre, N. Burgos, Reproducibility in Machine Learning for Medical Imaging, Springer US, New York, NY, 2023, pp. 631–653. doi:10.1007/978-1-0716-3195-9_21.
URL https://doi.org/10.1007/978-1-0716-3195-9_21

[9] J. S. Baxter, P. Jannin, Validation in the age of machine learning: A framework for describing validation with examples in transcranial magnetic stimulation and deep brain stimulation, Intelligence-Based Medicine 7 (2023) 100090. doi:https://doi.org/10.1016/j.ibmed.2023.100090.
URL https://www.sciencedirect.com/science/article/pii/S2666521223000042

[10] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M. A. Riegler, M. Wiesenfarth, A. E. Kavur, C. H. Sudre, M. Baumgartner, M. Eisenmann, D. Heckmann-NÃűtzel, A. T. RÃďdsch, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M. Blaschko, M. J. Cardoso, V. Cheplygina, B. A. Cimini, G. S. Collins, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, R. Haase, D. A. Hashimoto, M. M. Hoffman, M. Huisman, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, H. Kenngott, F. Kofler, A. Kopp-Schneider, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, K. G. M. Moons, H. MÃijller, B. Nichyporuk, F. Nickel, J. Petersen, N. Rajpoot, N. Rieke, J. Saez-Rodriguez, C. I. SÃąnchez, S. Shetty, M. van Smeden, R. M. Summers, A. A. Taha, A. Tiulpin, S. A. Tsaftaris, B. V. Calster, G. Varoquaux, P. F. JÃďger, Metrics reloaded: Recommendations for image analysis validation (2023). arXiv:2206.01653.

[11] S.-W. Lin, V. M. Bier, A study of expert overconfidence, Reliability Engineering & System Safety 93 (5) (2008) 711–721.

[12] C. L. Ramspek, K. J. Jager, F. W. Dekker, C. Zoccali, M. van Diepen, External valida-

tion of prognostic models: what,Âăwhy, how, when and where?, Clinical Kidney Journal 14 (1) (2020) 49–58. arXiv:https://academic.oup.com/ckj/article-pdf/14/1/49/36184810/sfaa188.pdf, doi:10.1093/ckj/sfaa188.
URL https://doi.org/10.1093/ckj/sfaa188

[13] R. A. Poldrack, C. I. Baker, J. Durnez, K. J. Gorgolewski, P. M. Matthews, M. R. Munafò, T. E. Nichols, J.-B. Poline, E. Vul, T. Yarkoni, Scanning the horizon: Towards transparent and reproducible neuroimaging research, Nature reviews. Neuroscience 18 (2) (2017) 115–126. doi:10.1038/nrn.2016.167.

[14] G. Varoquaux, O. Colliot, Evaluating machine learning models and their diagnostic value, in: O. Colliot (Ed.), Machine Learning for Brain Disorders, Springer, 2023.

[15] T. J. Bradshaw, Z. Huemann, J. Hu, A. Rahmim, A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging, Radiology: Artificial Intelligence 5 (4) (2023) e220232. doi:10.1148/ryai.220232.

[16] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, C. Ré, Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging, Proceedings of the ACM Conference on Health, Inference, and Learning 2020 (2020) 151–159. doi:10.1145/3368555.3384468.

[17] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, J. Santos, Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier], IEEE Computational Intelligence Magazine 13 (4) (2018) 59–76. doi:10.1109/MCI.2018.2866730.

[18] I. Tougui, A. Jilbab, J. El Mhamdi, Impact of the Choice of Cross-Validation Techniques on the Results of Machine Learning-Based Diagnostic Applications, Healthcare Informatics Research 27 (3) (2021) 189–199. doi:10.4258/hir.2021.27.3.189.

[19] J. M. Wicherts, C. L. S. Veldkamp, H. E. M. Augusteijn, M. Bakker, R. C. M. van Aert, M. A. L. M. van Assen, Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking, Frontiers in Psychology 7.

[20] T. E. Hardwicke, E.-J. Wagenmakers, Reducing bias, increasing transparency and calibrating confidence with preregistration, Nature Human Behaviour 7 (1) (2023) 15–26. doi:10.1038/s41562-022-01497-2.

[21] G. Varoquaux, V. Cheplygina, Machine learning for medical imaging: Methodological failures and recommendations for the future, npj Digital Medicine 5 (1) (2022) 1–8. doi:10.1038/s41746-022-00592-y.

[22] R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock, P. Avesani, B. M. Baczkowski, A. Bajracharya, L. Bakst, S. Ball, M. Barilari, N. Bault, D. Beaton, J. Beitner, R. G. Benoit, R. M. Berkers, J. P. Bhanji, B. B. Biswal, S. Bobadilla-Suarez, T. Bortolini, K. L. Bottenhorn, A. Bowring, S. Braem, H. R. Brooks, E. G. Brudner, C. B. Calderon, J. A. Camilleri, J. J. Castrellon, L. Cecchetti, E. C. Cieslik, Z. J. Cole, O. Collignon, R. W. Cox, W. A. Cunningham, S. Czoschke, K. Dadi, C. P. Davis, A. De Luca, M. R. Delgado, L. Demetriou, J. B. Dennison, X. Di, E. W. Dickie, E. Dobryakova, C. L. Donnat, J. Dukart, N. W. Duncan, J. Durnez, A. Eed, S. B. Eickhoff, A. Erhart, L. Fontanesi, G. M. Fricke, S. Fu, A. Galván, R. Gau, S. Genon, T. Glatard, E. Glerean, J. J. Goeman, S. A. E. Golowin, C. González-García, K. J. Gorgolewski, C. L. Grady, M. A. Green, J. F. Guassi Moreira, O. Guest, S. Hakimi, J. P. Hamilton, R. Hancock, G. Handjaras, B. B. Harry, C. Hawco, P. Herholz, G. Herman, S. Heunis, F. Hoffstaedter, J. Hogeveen, S. Holmes, C.-P. Hu, S. A. Huettel, M. E. Hughes, V. Iacovella, A. D. Iordan, P. M. Isager, A. I. Isik, A. Jahn, M. R. Johnson, T. Johnstone, M. J. E. Joseph, A. C. Juliano, J. W. Kable, M. Kassinopoulos, C. Koba, X.-Z. Kong, T. R. Koscik, N. E.

Kucukboyaci, B. A. Kuhl, S. Kupek, A. R. Laird, C. Lamm, R. Langner, N. Lauharatanahirun, H. Lee, S. Lee, A. Leemans, A. Leo, E. Lesage, F. Li, M. Y. Li, P. C. Lim, E. N. Lintz, S. W. Liphardt, A. B. Losecaat Vermeer, B. C. Love, M. L. Mack, N. Malpica, T. Marins, C. Maumet, K. McDonald, J. T. McGuire, H. Melero, A. S. Méndez Leal, B. Meyer, K. N. Meyer, G. Mihai, G. D. Mitsis, J. Moll, D. M. Nielson, G. Nilsonne, M. P. Notter, E. Olivetti, A. I. Onicas, P. Papale, K. R. Patil, J. E. Peelle, A. Pérez, D. Pischedda, J.-B. Poline, Y. Prystauka, S. Ray, P. A. Reuter-Lorenz, R. C. Reynolds, E. Ricciardi, J. R. Rieck, A. M. Rodriguez-Thompson, A. Romyn, T. Salo, G. R. Samanez-Larkin, E. Sanz-Morales, M. L. Schlichting, D. H. Schultz, Q. Shen, M. A. Sheridan, J. A. Silvers, K. Skagerlund, A. Smith, D. V. Smith, P. Sokol-Hessner, S. R. Steinkamp, S. M. Tashjian, B. Thirion, J. N. Thorp, G. Tinghög, L. Tisdall, S. H. Tompson, C. Toro-Serey, J. J. T. Tresols, L. Tozzi, V. Truong, L. Turella, A. E. van 't Veer, T. Verguts, J. M. Vettel, S. Vijayarajah, K. Vo, M. B. Wall, W. D. Weeda, S. Weis, D. J. White, D. Wisniewski, A. Xifra-Porxas, E. A. Yearling, S. Yoon, R. Yuan, K. S. Yuen, L. Zhang, X. Zhang, J. E. Zosky, T. E. Nichols, R. A. Poldrack, T. Schonberg, Variability in the analysis of a single neuroimaging dataset by many teams, Nature 582 (7810) (2020) 84–88. doi:10.1038/s41586-020-2314-9.

[23] J. R. Baldwin, J.-B. Pingault, T. Schoeler, H. M. Sallis, M. R. Munafò, Protecting against researcher bias in secondary data analysis: Challenges and potential solutions, European Journal of Epidemiology 37 (1) (2022) 1–10. doi:10.1007/s10654-021-00839-0.

[24] T. E. Hardwicke, J. P. A. Ioannidis, Mapping the universe of registered reports, Nature Human Behaviour 2 (11) (2018) 793–796. doi:10.1038/s41562-018-0444-y.

[25] A. Rule, A. Birmingham, C. Zuniga, I. Altintas, S.-C. Huang, R. Knight, N. Moshiri, M. H. Nguyen, S. B. Rosenthal, F. Pérez, P. W. Rose, Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks, PLOS Computational Biology 15 (7) (25 juil. 2019) e1007007. doi:10.1371/journal.pcbi.1007007.

[26] T. Glatard, L. B. Lewis, R. Ferreira da Silva, R. Adalat, N. Beck, C. Lepage, P. Rioux, M.-E. Rousseau, T. Sherif, E. Deelman, N. Khalili-Mahani, A. C. Evans, Reproducibility of neuroimaging analyses across operating systems, Frontiers in Neuroinformatics 9. doi:10.3389/fninf.2015.00012.
URL https://www.frontiersin.org/articles/10.3389/fninf.2015.00012

[27] K. Gorgolewski, C. Burns, C. Madison, D. Clark, Y. Halchenko, M. Waskom, S. Ghosh, Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python, Frontiers in Neuroinformatics 5. doi:10.3389/fninf.2011.00013.
URL https://www.frontiersin.org/articles/10.3389/fninf.2011.00013

[28] S. Cohen-Boulakia, K. Belhajjame, O. Collin, J. Chopard, C. Froidevaux, A. Gaignard, K. Hinsen, P. Larmande, Y. L. Bras, F. Lemoine, F. Mareuil, H. MÃľnager, C. Pradal, C. Blanchet, Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities, Future Generation Computer Systems 75 (2017) 284–298. doi:https://doi.org/10.1016/j.future.2017.01.012.
URL https://www.sciencedirect.com/science/article/pii/S0167739X17300316

[29] D. B. Stockton, A. A. Prinz, F. Santamaria, Provenance and reproducibility in the automation of a standard computational neuroscience pipeline, in: Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems, P-RECS '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 7âĂŞ12. doi:10.1145/3322790.3330592.
URL https://doi.org/10.1145/3322790.3330592

[30] G. Wilson, D. A. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, P. Wilson,

Best practices for scientific computing, PLOS Biology 12 (1) (2014) 1–7. doi:10.1371/journal.pbio.1001745.
URL https://doi.org/10.1371/journal.pbio.1001745

[31] B. Fischl, Freesurfer, Neuroimage 62 (2) (2012) 774–781.

[32] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, S. M. Smith, Fsl, NeuroImage 62 (2) (2012) 782–790, 20 YEARS OF fMRI. doi:https://doi.org/10.1016/j.neuroimage.2011.09.015.
URL https://www.sciencedirect.com/science/article/pii/S1053811911010603

[33] D. N. Kennedy, S. A. Abraham, J. F. Bates, A. Crowley, S. Ghosh, T. Gillespie, M. Goncalves, J. S. Grethe, Y. O. Halchenko, M. Hanke, et al., Everything matters: the repronim perspective on reproducible neuroimaging, Frontiers in neuroinformatics (2019) 1.

[34] S. J. Eglen, B. Marwick, Y. O. Halchenko, M. Hanke, S. Sufi, P. Gleeson, R. A. Silver, A. P. Davison, L. Lanyon, M. Abrams, et al., Toward standard practices for sharing computer code and programs in neuroscience, Nature neuroscience 20 (6) (2017) 770–773.

[35] A. Bowring, C. Maumet, T. E. Nichols, Exploring the impact of analysis software on task fmri results, Human brain mapping 40 (11) (2019) 3362–3384.

[36] Y. O. Halchenko, M. Hanke, Open is not enough. let's take the next step: an integrated, community-driven computing platform for neuroscience, Frontiers in neuroinformatics 6 (2012) 22.

[37] N. Vallet, D. Michonneau, S. Tournier, Toward practical transparent verifiable and long-term reproducible research using guix, Scientific Data.
URL https://doi.org/10.1038/s41597-022-01720-9

[38] M. Hatt, B. Laurent, A. Ouahabi, H. Fayad, S. Tan, L. Li, W. Lu, V. Jaouen, C. Tauber, J. Czakon, F. Drapejkowski, W. Dyrka, S. Camarasu-Pop, F. Cervenansky, P. Girard, T. Glatard, M. Kain, Y. Yao, C. Barillot, A. Kirov, D. Visvikis, The first MICCAI challenge on PET tumor segmentation, Medical Image Analysis 44 (2018) 177–195. doi:10.1016/j.media.2017.12.007.
URL https://hal.science/hal-01659162

[39] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. C. Pop, P. Girard, R. Ameli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, T. Glatard, J. Beaumont, S. Doyle, F. Forbes, J. Knight, A. Khademi, A. Mahbod, C. Wang, R. Mckinley, F. Wagner, J. Muschelli, E. Sweeney, E. Roura, X. Llado, M. Santos, W. P. Santos, A. G. Silva-Filho, X. Tomas-Fernandez, H. Urien, I. Bloch, S. Valverde, M. Cabezas, F. J. Vera-Olmos, N. Malpica, C. R. G. Guttmann, S. Vukusic, G. Edan, M. Dojat, M. Styner, S. K. Warfield, F. Cotton, C. Barillot, Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure, Scientific Reports 8 (1) (2018) 13650. doi:10.1038/s41598-018-31911-7.
URL https://inserm.hal.science/inserm-01847873

[40] L. Maier-Hein, A. Reinke, M. Kozubek, A. L. Martel, T. Arbel, M. Eisenmann, A. Hanbury, P. Jannin, H. Müller, S. Onogur, et al., Bias: Transparent reporting of biomedical image analysis challenges, Medical image analysis 66 (2020) 101796.

[41] M. Canesche, R. Leissa, F. M. Q. Pereira, Preparing reproducible scientific artifacts using docker, arXiv preprint arXiv:2308.14122.

[42] L. H. Boulogne, J. Lorenz, D. Kienzle, R. Schon, K. Ludwig, R. Lienhart, S. Jegou, G. Li, C. Chen, Q. Wang, et al., The stoic2021 covid-19 ai challenge: applying reusable training methodologies to private data, arXiv preprint arXiv:2306.10484.

[43] E. Bannier, G. Barker, V. Borghesani, N. Broeckx, P. Clement, K. Emblem, S. Ghosh, E. Glerean, K. Gorgolewski, M. Havu, Y. Halchenko, P. Herholz, A. Hespel, S. Heunis, Y. Hu,

C. Hu, D. Huijser, M. de la Iglesia VayÃą, R. Jancalek, V. Katsaros, M. Kieseler, C. Maumet, C. Moreau, H. Mutsaerts, R. Oostenveld, E. Ozturk-Isik, N. Pascual Leone Espinosa, J. Pellman, C. Pernet, F. Pizzini, A. TrbaliÄĞ, P. Toussaint, M. Visconti di Oleggio Castello, F. Wang, C. Wang, H. Zhu, The open brain consent: Informing research participants and obtaining consent to share brain imaging data., Hum Brain Mapp.doi:10.1002/hbm.25351.

[44] M. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, L. da Silva Santos, P. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, A. Gonzalez-Beltran, A. Gray, P. Groth, C. Goble, J. Grethe, J. Heringa, P. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. Lusher, M. Martone, A. Mons, A. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, P. Waagmeester, A ans Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The fair guiding principles for scientific data management and stewardship., Sci Datadoi:10.1038/sdata.2016.18.

[45] C. Barillot, E. Bannier, O. Commowick, I. Corouge, A. Baire, I. Fakhfakh, J. Guillaumont, Y. Yao, M. Kain, Shanoir: Applying the software as a service distribution model to manage brain imaging research repositories, Frontiers in ICT 3. doi:10.3389/fict.2016.00025.
URL https://www.frontiersin.org/articles/10.3389/fict.2016.00025

[46] S. Das, A. P. Zijdenbos, J. Harlap, D. Vins, A. C. Evans, Loris: a web-based data management system for multi-center studies, Front Neuroinformdoi:10.3389/fninf.2011.00037.

[47] Y. O. Halchenko, K. Meyer, B. Poldrack, D. S. Solanky, A. S. Wagner, J. Gors, D. MacFarlane, D. Pustina, V. Sochat, S. S. Ghosh, C. MÃűnch, C. J. Markiewicz, L. Waite, I. Shlyakhter, A. de la Vega, S. Hayashi, C. O. HÃđusler, J.-B. Poline, T. Kadelka, K. SkytÃľn, D. Jarecka, D. Kennedy, T. Strauss, M. Cieslak, P. Vavra, H.-I. Ioanas, R. Schneider, M. PflÃijger, J. V. Haxby, S. B. Eickhoff, M. Hanke, Datalad: distributed system for joint management of code, data, and their relationship, Journal of Open Source Software 6 (63) (2021) 3262. doi:10.21105/joss.03262.
URL https://doi.org/10.21105/joss.03262

[48] S. Noble, D. Scheinost, R. T. Constable, A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis, Neuroimage 203 (2019) 116157.

[49] A. Plant, R. Hanisch, Reproducibility in Science: A Metrology Perspective, Harvard Data Science Review 2 (4), https://hdsr.mitpress.mit.edu/pub/0r4v4k4z.

[50] A. Mason, J. Rioux, S. E. Clarke, A. Costa, M. Schmidt, V. Keough, T. Huynh, S. Beyea, Comparison of objective image quality metrics to expert radiologistsâĂŹ scoring of diagnostic quality of mr images, IEEE transactions on medical imaging 39 (4) (2019) 1064–1072.

[51] A. Naëgel, H. Ratiney, J. Karkouri, D. Kennouche, N. Royer, J. M. Slade, J. Morel, P. Croisille, M. Viallon, Alteration of skeletal muscle energy metabolism assessed by phosphorus-31 magnetic resonance spectroscopy in clinical routine, part 1: Advanced quality control pipeline, NMR in Biomedicine (2023) e5025.

[52] W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical bayes methods, Biostatistics 8 (1) (2007) 118–127.

[53] F. Orlhac, J. J. Eertink, A.-S. Cottereau, J. M. Zijlstra, C. Thieblemont, M. Meignan, R. Boellaard, I. Buvat, A guide to combat harmonization of imaging biomarkers in multicenter studies, Journal of Nuclear Medicine 63 (2) (2022) 172–179. arXiv:https://jnm.snmjournals.org/content/63/2/172.full.pdf, doi:10.2967/jnumed.121.262464.
URL https://jnm.snmjournals.org/content/63/2/172

[54] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, et al., The future of digital health with federated learning, NPJ

digital medicine 3 (1) (2020) 119.