Chapter 21

Reproducibility in Machine Learning for Medical Imaging

Olivier Colliot, Elina Thibeau-Sutre, and Ninon Burgos

Abstract

Reproducibility is a cornerstone of science, as the replication of findings is the process through which they become knowledge. It is widely considered that many fields of science are undergoing a reproducibility crisis. This has led to the publications of various guidelines in order to improve research reproducibility.

This didactic chapter intends at being an introduction to reproducibility for researchers in the field of machine learning for medical imaging. We first distinguish between different types of reproducibility. For each of them, we aim at defining it, at describing the requirements to achieve it, and at discussing its utility. The chapter ends with a discussion on the benefits of reproducibility and with a plea for a nondogmatic approach to this concept and its implementation in research practice.

Key words Reproducibility, Replicability, Reliability, Repeatability, Open science, Machine learning, Artificial intelligence, Deep learning, Medical imaging

1 Introduction

Reproducibility is at the core of the scientific method. In its general and most common meaning, it corresponds to the ability to reproduce the findings of a given experimental study. This is a necessary (but not sufficient) condition for a scientific statement to become accepted as new knowledge. Let's illustrate this with a simple example, considering the following statement: "the volume of the hippocampus is, on average, smaller in patients with Alzheimer's disease (AD) than in healthy people of comparable age." Such statement was the conclusion of studies which measured such volume from magnetic resonance images (MRI). To the best of our knowledge, the first study to assert this was that of Seab et al [1]. This was later reproduced by many other studies (e.g., [2, 3]). It is now widely accepted, which would not have been the case if the study had proven impossible to reproduce. Note that, as stated above, this is a *necessary* but not a *sufficient* condition. Indeed, there could be other reasons for such statement not to be considered as knowledge. For instance, let's imagine that some other

Olivier Colliot (ed.), *Machine Learning for Brain Disorders*, Neuromethods, vol. 197, https://doi.org/10.1007/978-1-0716-3195-9_21,

researchers discover that there is an artifact that is systematically present in the MRI of patients with AD and which leads to erroneous volume estimation. Then, the statement could not be considered new knowledge even though it had been reproduced several times.

Machine learning (ML) is, in part, an experimental science. This is not the case of the entirety of the discipline, part of which is theoretical (for instance, mathematical proofs of convergence or of approximation capabilities of different classes of models) or methodological (the invention of a new approach). Nevertheless, since ML ultimately aims at solving practical problems, its experimental component is essential. Typically, one would want to be able to make statements of the type described above from an experimental study. Here is an example of such statement: "this ML model (for instance, a specific convolutional neural network [CNN] architecture), using MRI data as input, is capable of classifying AD patients and healthy controls with an accuracy superior to 80%." In order to end an article with such a statement, one needs to conduct an experimental study. For such findings to become knowledge, it needs to be subsequently reproduced. Of course, this statement is unlikely to be universal, and one would want to know under which conditions it holds: for instance, is it restricted to a specific class of MRI scanners, to specific disease stages, to specific age ranges?

Box 1: Glossary
The readers will find the definition of the terms we used in the present document.

- **Reproducibility, replicability, repeatability.** In the present document, these will be used as synonyms of reproducibility.
- **Original study.** Study that first showed a finding.
- **Replication study.** Study that subsequently aimed at replicating an original study, with the hope to support its findings.
- **Research artifact.** Any output of scientific research: papers, code, data, protocols.... Not to be confused with imaging artifacts which are defects of imaging data.
- **Claims.** The conclusions of a study. Basically a set of statements describing the results and a set of limitations which delineate the boundaries within which the claims are stated (the term "claim" is here used in the broad scientific sense not with the specific meaning it has in the context of regulation of medical devices although the two may be related).
- **Limitations.** A set of restrictions under which the claims may not hold (usually because the corresponding settings have not been explored).

(continued)

Box 1 (continued)

- **Method.** The ML approach described in the paper, independently of its implementation.
- **Code.** The implementation of the method.
- **Software dependencies.** Other software packages that the main code relies on and which are necessary for its execution.
- **Public data.** Data that can be accessed by anybody with no or little restriction (for instance, the data hosted at https://openneuro.org).
- **Semi-public data.** Data which requires approval of a research project (for instance, the Alzheimer's Disease Neuroimaging Initiative [ADNI] http://www.adni-info.org). The researchers can then use the data only for the intended research purpose and cannot redistribute it.
- **Data split.** Separation into training, validation, and test sets.
- **Data leakage.** Faulty procedure which has led information from the training set to leak into the test set. *See* refs. 4, 5 for details.
- **Error margins.** A general term for providing the precision of the performance estimates (e.g., standard-error or confidence intervals).
- **Researcher degrees of freedom.** Number of different components (e.g., different architectures, hyperparameter values, subsamples...) which have been tried before arriving to the final method [6]. Too many degrees of freedom tend to produce methods that do not generalize.
- **p-hacking.** A bad practice that involves too many degrees of freedom and which consists is trying many different statistical procedures until a significant p-value is found.
- **Acquisition settings.** Factors that influence the scan of a given patient (imaging device, acquisition paratemeters, image quality).
- **Image artifacts.** Defects of a medical image, these may include noise, field heterogeneity, motion artifacts, and others.
- **Preregistration.** The deposit of the study protocol prior to performing the study. Limits degrees of freedom and increases likelihood of robust findings.

In the examples above, we have actually illustrated only one on the many possible meanings of reproducibility: the addition of new evidence to support a scientific finding of an *original* study through reproduction under different experimental conditions (*see* Box 1 for a glossary of some of the key concepts used). However, it is also used for very different meanings. In computational sciences, it is often used for the ability to exactly reproduce the results (i.e., the exact numbers) in a given study. In sciences which aim at providing measurements (as is often the case in medical imaging), the word may be used to describe the variability of a given measurement tool under different acquisition settings. We shall provide more details on these different meanings in Subheading 2. Finally, the topics of reproducibility and open science are obviously related since the latter favors the former. However, open science encompasses a broader objective which is to make all *research artifacts* (code, data, papers. . .) openly available for the benefit of the whole society. Conversely, open research may still be unreproducible (e.g., because it has relied on faulty statistical procedures).

There has been increasing concern that science is undergoing a reproducibility crisis [7–10]. This is present in various fields from psychology [11] to preclinical oncology research [12]. ML [13–17], digital medicine [18], ML for healthcare [19, 20], and ML for medical imaging [21] are no exception. The concerns are multifaceted. In particular, they include two substantially different aspects: the report of failures to reproduce previous studies and the observation that many papers do not provide sufficient information for reproducing their results. It is important to have in mind that, while the two may be related, there is not a direct relationship between them: it may very well be that a paper seems to include all the necessary information for reproduction and that reproduction attempts fail (for instance, because the original study had too many degrees of freedom and led to a method that only works on a single dataset, *see* Subheading 4).

Various guidelines have been proposed to improve research reproducibility. Such guidelines may be general [10] or devoted to specific fields including brain imaging [22–25] and ML for healthcare and life sciences [26, 27]. Moreover many other papers, even though not strictly providing guidelines, provide very valuable pieces of advice for making research more trustworthy and in particular more reproducible (e.g., [14, 19, 28–32]).

This chapter is an introduction to the topic of reproducibility for researchers in the field of ML for medical imaging. It is not meant at providing a replacement for the aforementioned previously published guidelines that we strongly encourage the reader to refer to.

The remainder of the chapter is organized as follows. We first start by introducing different types of reproducibility (Subheading 2). For each of them, we attempt to clearly define it and describe what are the requirements to achieve it and the benefits it can provide (Subheadings 3, 4, 5, and 6). All this information is given with having the field of ML for medical imaging as a target, even though part of it may apply to other fields. Finally, we conclude with a discussion which both describes the benefits of reproducibility but also advocates for a nondogmatic point of view on the topic (Subheading 7).

2 The Polysemy of Reproducibility

The term "reproducibility" has been used with various meanings which may range from the exact reproduction of a study with the same material and methods, to the reproduction of a result using new experimental data to the support of a scientific idea using a completely different experimental setup [33, 34]. Moreover, various terms have been introduced including reproducibility, replicability, repeatability, reliability, robustness, generalizability. . .Some of these words, for instance, reproducibility vs replicability, have even been used by some authors with opposite meanings [33, 34]. We will not aim at assigning an unambiguous meaning to each of these words, as we find this of little interest, and will use the term "reproducibility," "replicability," and "repeatability" as synonyms. On the other hand, we believe, as many other authors [19, 23, 33, 35], that it is important to distinguish between different types of reproducibility. To that purpose, it is useful to have a *taxonomy* of reproducibility. Below, we describe such a taxonomy. We do not claim that it is novel, as it takes inspiration from other papers [14, 19, 23, 33, 35] nor that it should be universally adopted. Furthermore, boundaries between different types of reproducibility are partly fuzzy. We simply hope that it will be useful for the different concepts that we subsequently introduce and that it will be well adapted to the field of ML for medical imaging.

We distinguish between four main types of reproducibility: **exact reproducibility**, **statistical reproducibility**, **conceptual reproducibility**, and **measurement reproducibility**. We describe those four main types in the following sections. They are also summarized in Fig. 1. As will be explained below, the three first types have relationships with each other (this is why they have the same color in the figure) while the fourth is more separated.

3 Exact Reproducibility

What Is It? Exact reproducibility aims at reproducing strictly identical results as those of a previously published paper. Concretely, this amounts to being able to reproduce tables and figures as they appear in the original paper following the same procedures as the authors.

What Does It Require? Exact reproducibility requires to have access to all components that led to the results including, of course, data and code.

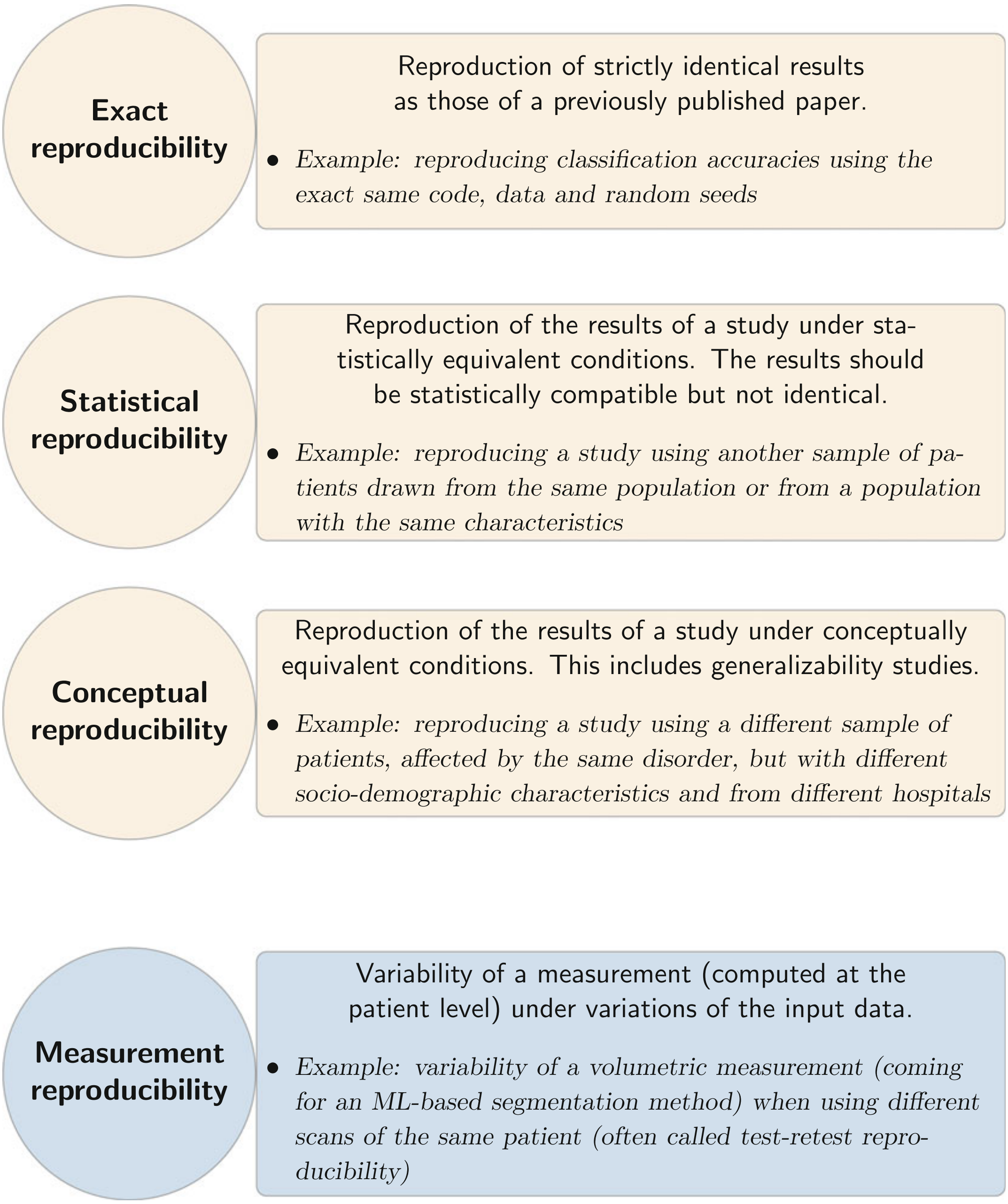

Fig. 1 Different types of reproducibility. Note that, in the case of "statistical" and "conceptual" reproducibility, the terms come from [19] but the exact definition provided in each corresponding section may differ

Access to data is obviously necessary [19, 22, 27]. Open data has been described (together with code and papers) as one of the pillars of open science [22]. It is widely accepted that scientific data should adhere to the FAIR (Findable, Accessible, Interoperable, Reusable) principles (please refer to https://www.go-fair.org/fair-principles and [36] for more details). Among these principles, accessibility is often the most difficult to adhere to for medical imaging data (or healthcare data in general). It is very common in medical papers that data is mentioned as available upon request. However, a study has showed that, when data is subsequently requested, many researchers actually do not comply with the data accessibility statement [37]. This is worrisome, and more transparent ways of data sharing would be welcome. However, as mentioned above, such transparent sharing procedures may be difficult to put in place for healthcare data. In particular, making the data public is often difficult due to regulatory and privacy constraints [19]. Gorgolewski and Poldrack [22] provide useful pieces of advice to facilitate sharing, but there are cases where public sharing will remain impossible. In particular, one must distinguish between research data (acquired as part of a research protocol), which can often be made *public* or *semi-public*[1] provided that adequate measures have been taken at data collection (e.g., adequate participant consent), and routine clinical data (acquired as part of the routine clinical care of the patients), which sharing can be much more complicated. It is important that data is easily findable and that it is shared on a server which has a long-term maintenance. General purpose data repositories such as Zenodo[2] provide a good solution. Another important aspect is to adhere to community standards for data organization, so that it can easily be reused by researchers. For brain imaging, the community standard is BIDS (Brain Imaging Data Structure) [38].[3] This standard is already very mature and has already been extended to incorporate other modalities such as microscopy images, for instance (Microscopy-BIDS [39]). Note that there is an ongoing proposal to extend it to other organs (MIDS – Medical Imaging Data Structure [40][4]). Finally, we would like to draw the attention to an important point that is often overlooked. Even when a study relies on public or semi-public data, it is absolutely necessary to specify which samples (e.g., which participants and which scans) have been used; otherwise, the study is not reproducible [41]. Ideally, one would provide code to automatically make the data selection [42] in order to ease the replication.

Another key component is that the code is accessible [19, 22, 27]. Indeed, it would be delusional to think that exactly the same results could be obtained using a reimplementation based on information provided in the paper (even though it is good practice to provide as much detail as possible about the methods in the paper).

[1] *See* glossary Box 1.

[2] https://zenodo.org/.

[3] https://bids.neuroimaging.io/.

[4] See BIDS extension proposal (BEP) number 25 (BEP025) https://bids.neuroimaging.io/get_involved.html#extending-the-bids-specification

Theoretically, it does not mean that the code must come with an open software license. However, doing so has many additional benefits such as allowing other researchers to use the code or parts of it for different purposes. The code should be made according to good coding practices which include the use of a versioning system and adequate documentation [20]. Furthermore, although not strictly needed for reproducibility, the use of continuous integration makes the code more robust and eases its long-term maintenance. Besides, it is good to ease as much as possible the installation of dependencies [27]. This can be done with pip[5] when programming in Python. One can also use containers such as Docker.[6] One can find useful additional advice in the *Tips for Publishing Research Code*.[7] Note that we are not saying that all these components should be present in any study or are prerequisites for good research. They constitute an ideal reproducibility goal.

Sharing well-curated notebooks is also a way to ease reproducibility of results by other researchers. This can be done through standard means, but dedicated servers also exist. One can, for instance, cite an interesting initiative called NeuroLibre[8] which provides a preprint server for reproducible data analysis in neuroscience, in particular providing curated and reviewed Jupyter notebooks [43].

In ML, sharing the code itself is not enough for exact reproducibility. First, every element of the training procedure should be stored: this includes the data splits and the criteria for model selection. Moreover, there usually are non-deterministic components so it is necessary to store random seeds [27]. Furthermore, software/operating system versions, the GPU model/version, and threading have been deemed necessary to obtain exact reproducibility [44]. The ClinicaDL software platform provides a framework for easing exact reproducibility of deep learning for neuroimaging [5].[9] Although it is targeted at brain imaging, many of its components and concepts are applicable to medical imaging in general. Also in the field of brain imaging, NiLearn[10] facilitates the reproducibility of statistics and ML. One can also cite Pymia[11] which provides data handling and validation functionalities for deep learning in medical imaging [45].

[5] https://pypi.org/project/pip/.

[6] https://www.docker.com/.

[7] https://github.com/paperswithcode/releasing-research-code.

[8] https://neurolibre.org/.

[9] https://clinicadl.readthedocs.io/en/latest/.

[10] https://nilearn.github.io/stable/index.html.

[11] https://github.com/rundherum/pymia.

Finally, it may seem obvious, but, even when the code is shared, the underlying theory of the method, all its components, and implementation details need to be present in the paper [14].

It is in principle possible to retrain models identically if the above elements are provided. It nevertheless remains a good practice to share trained models, in order to allow other researchers to check that retraining indeed led to the same results but also to save computational resources. However, models can be attacked to recover training data [27, 46]. This is not a problem when the training data is public. When it is privacy-sensitive, methods to preserve privacy exist [27, 47].

In medical imaging, preprocessing and feature extraction are often critical steps that will subsequently influence the ML results. It is thus necessary to also provide code for such parts. Several software initiatives including BIDSApps [48][12] and Clinica [49][13] provide ready-to-use tools for preprocessing and feature extraction for various brain imaging modalities. Applicable to many medical imaging modalities, the ITK [50, 51][14] framework provides a wide range of processing tools. It can ease the work of researchers who do not want to spend time on preprocessing and feature extraction pipelines and focus on the ML part of their work.

Why Is It Useful? It has been claimed that exact reproducibility is of little interest, that pursuing it is a waste of energy of the community, and that its only possible use would be the detection of outright fraud which is rare [52]. We disagree with that view. Let's start with fraud. It may be of low occurrence, although its exact prevalence is difficult to establish. Even so, it is of disastrous consequences as it leads to loss of trust by students, scientists, and the general public. In particular, a survey of 1,576 researchers indicated that 40% of them believe that fraud is a factor that "always/often" contributes to irreproducible research and that 70% of them think that it "sometimes" contributes [7]. Exact reproducibility can certainly contribute to reduce fraud as full transparency obviously makes fraud more difficult. Fraud remains possible (one could forge some data and share it), but it is more difficult to achieve under transparency constraints. Fraud may be rare but errors are much more common. The framework of exact reproducibility eases the detection of errors which is a service to science and even to the authors themselves. In particular, it may help discover "biases and artifacts in the data that were missed by the authors and that cannot be discovered if the data are never made available" [27]. Similarly, it can lead to the discovery of

[12] https://bids-apps.neuroimaging.io/.
[13] https://www.clinica.run/.
[14] https://itk.org/.

wrong validation schemes, including data leakage or errors in implementation that make it inconsistent with the methodology presented in the paper. Overall, it may make progress slower, but it will definitely make it steadier. However, this does not mean that exact reproducibility should be aimed in all works or made a requirement for all publications (*see* Subheading 7.4).

4 Statistical Reproducibility

What Is It? Statistical reproducibility aims at reproducing findings under statistically equivalent conditions.[15] The specific definition may vary, but the following choices are often considered reasonable. The implementation of the method (the code) remains the same. Random components are left random. Regarding the data, the general idea is that the sample would be drawn from the same population. One could, for instance, use subsamples of the original data or another subsample of a larger source population. An interesting case is to study different data splits. A less restrictive view of statistical reproducibility would be to use another dataset whose characteristics are similar to those of the original dataset (for instance, similar age, sex, scanner distributions). Note that the boundary between statistical and conceptual reproducibility (defined in the next section) is fuzzy. We do not believe it is possible to draw exact frontiers that would delimit the statistical variations that are admissible in a statistical reproducibility study. Finally, it is important that those who conduct the statistical replication study clearly indicate which components of variability they study.

What Does It Require? Here one needs to distinguish between two types of factors: those necessary to *attempt* reproducibility and those that increase the likelihood of *successful* reproducibility.

Regarding the first type, most factors are common with those for exact reproducibility. Code needs to be accessible so that variations coming from reimplementation do not impact the replication. Random seeds, GPU model, or other software/execution parameters will not be set to be identical because the aim is precisely to check if the findings of the study are statistically reproducible under such variations. Knowing their value in the original study is nevertheless useful in order to dissect potential reasons for failed replication. Trained models are in a similar situation: they will usually not be used for statistical replication (models will be retrained) but shall prove useful to dissect potential failures. Data

[15] We use the term of [19] although with a slightly different (more extensive) meaning.

accessibility is also very valuable because it will allow studying different data splits, or subsamples.

The abovementioned elements make it possible for other researchers to attempt statistical replication of a given study. On the other hand, there are features of the original study that will make such replication more likely to be successful (equivalently, one could say that the original findings are robust). One important factor is that the original study reports error margins (reporting the standard error or equivalently a confidence interval). It is important in this specific context because statistical reproducibility does not aim at obtaining (and cannot obtain) exactly the same results. One wants the results to be *compatible* with original ones: typically a successful replication would produce results which are within the error margin of the original study. Beyond the topic of statistical reproducibility, the report of error margins is of great importance in general, in particular in the field of ML for medical imaging, because it provides a precision on the estimates of the performance. Unfortunately, this practice is still too uncommon in the ML field as a whole [19]. Even worse, it is not uncommon to find faulty interpretations of estimates. For instance, one should never estimate standard errors (SE) from multiple runs of a cross-validation, as the number of runs can be made arbitrarily large and as a consequence the SE arbitrarily small (*see* [4]). A very common example is papers which report empirical standard deviation (SD) across *k*-folds (or more generally across splits). Unlike what is quite widely believed, this value does not allow to gauge the precision of the performance estimation. It provides some insight on the variability of the learning procedure under variations of the training and validation sets. Further, keep in mind that when *the number of splits* is small, such gauge will be very rough. *When the number of splits is* sufficiently large (and typically using random splits rather than *k*-fold), it is possible to assess if a "learner" (i.e., an ML procedure to perform a task) is superior to another one by counting the fraction of folds on which it obtains superior performance (e.g., 75%) [53]. *See* Chap. 21 for more details on this question. However, in no case can such procedures estimate the precision of the performance of the trained model, in other words the precision of the computed biomarker or computer-aided diagnosis tool. This requires an independent test set, from which SE and confidence intervals can be computed.

Why Is It Useful? Statistical replication has many merits. First, by reassessing ML methods using different data splits, one can spot faulty procedures including data leakage which is prevalent in the field of medical imaging [54–57]. *See* refs. 4, 5 for more details on data leakage. Beyond procedures which are clearly wrong, it can also detect lack of robustness to different parameters. One would consider that the procedure is not statistically replicable if it leads to substantially different results under different train/test data splits, different random seeds, or small changes in hyperparameters. Such an ML algorithm would display poor robustness and would be unlikely to be of future clinical use. Note that, regarding the use of different train/test data splits, these would need to preserve a distribution of metadata (for instance, age, sex, diagnosis...) between train and test that is similar to that of the original study. Most classically, if the original study has stratified the splits, the statistical replication study would also need to stratify the splits. Using different distributions (e.g., not stratified) is also interesting but, in our view, falls within conceptual rather than statistical reproducibility. Furthermore, it is very interesting to attempt replication on a different dataset with statistically equivalent characteristics: for instance, another subsample which has not been used in the original study (but comes from the same larger dataset) or a different dataset but with similar characteristics (e.g., same MRI scanners, similar age, similar disease stage...). Unsuccessful replication may be an indication of overfitting of the dataset of the original study through excessive experimentation with different architectures or hyperparameters which ended up with a method that would work only on this very specific dataset. This is referred to as the *researcher degrees of freedom* [6, 22]. This concept extends beyond the field of ML. It actually comes from experimental sciences where different statistical procedures are tried until a statistically significant result is found, a bad practice known as *p-hacking* [58]. It is important that researchers in our field have this problem in mind. Experimental sciences have proposed *preregistered* and *registered* studies as a potential solution to ban such bad practices. Preregistration means that the research plan is written down and made public before the study starts. It can, for example, be published on the Open Science Framework website.[16] This mechanism reduces the researcher degrees of freedom and is thus likely to lead to more robust results. Registration goes one step further. The research plan is submitted to a journal and peer-reviewed. Thus (most of) the peer review is done before the results are known. It has the additional advantage of putting more focus on methodological soundness than on the groundbreaking nature of results (for instance, negative results will be published). More details about preregistration and registration can be found in [59]. Preregistration and registration are not yet widely used in ML for medical imaging. Such practices would certainly not fit all studies because they leave no room for methodological creativity. On the other hand, they should be very valuable to experimental

[16] https://osf.io.

studies aiming at validating ML methods. We believe that, as a community, we should try to adapt such procedures to our field.

5 Conceptual Reproducibility

What Is It? Conceptual reproducibility can be seen as the ultimate goal: the one which lead to the consolidation of scientific knowledge. The general idea is to be able to validate the findings under conceptually similar conditions.[17] Conceptually similar means that the method, the data, and the experiments are compatible with the claims of the original study but they are not identical. We will come back to the notion of claims of a study, and their relationships to generalizability and limitations, later in this section.

What Does It Require? Again, we may distinguish between factors that make it possible to *attempt* replication and those that will make it more likely to be *successful*.

In theory, nothing but the original paper should be strictly necessary. Nevertheless, this assumes that the original paper has adhered to the scientific gold standard of providing all details necessary for replication: not only a description of the methods which makes reimplementation possible but a detailed description of the datasets and experimental procedure. It is particularly worrisome that many medical imaging publications do not even report basic demographic statistics [30]. [14] argues that the replication should be independent of the implementation. We agree in principle but believe that such requirement would considerably lower the number of conceptual replication attempts, while more are needed to advance our field in a steadier manner. In practice, it is extremely useful to be able to access the code, not only to save a lot of time but also to make sure that an unsuccessful replication is not due to a faulty reimplementation. The same can be said for trained models. Access to the original data can be useful to dissect the potential reasons for differences in results. In summary, none of the elements of exact reproducibility are required, all of them are welcome.

There are several characteristics of an original study that make it less likely for it to be replicated. Low sample size not only means that it is less likely to find a true effect if it exists but also increases the odds that a positive finding is false [9]. This is not only true in ML but in experimental sciences in general. Ideally, the sample size should be justified by a previous power analysis [24]. Causes for failure of statistical reproducibility also apply here. In particular, too many researcher degrees of freedom increase the likelihood of

[17] Again, we use the term of [19] although with a slightly different meaning.

having built a method that is overly specific to a dataset. Another problem is that the datasets used in medical imaging ML papers are very often not representative of what would be found in the clinic [30]. Indeed, they often come from research datasets where the inclusion criteria are specific, the medical imaging protocols are harmonized, and the data quality is controlled. Thus, it is necessary to have more studies including clinical routine data (e.g., [60, 61]). Finally, it is very important to have in mind that most scientific findings will not universally replicate but that the replication will only succeed under specific conditions. This is why it is critical that scientific papers precisely define their claims and their limitations. For instance, a claim could be that a given algorithm can segment brain tumors with a Dice of 0.9 ± 0.02 when the MR images are acquired at 3 Tesla and have only minimal artifacts. The same paper would mention as limitations that it is unclear how the algorithm would perform at 1.5 Tesla or with data of lower quality. One can see that stating clear claims and limitations will allow defining the scope of conceptual replication studies. Studies outside that scope would aim at studying generalizability beyond original claims.

Why Is It Useful? As mentioned above, conceptual reproducibility is the ultimate goal, the one which, through accumulation of evidence, builds consensus about new scientific knowledge. Its utility in general is thus obvious. More specifically, it provides different benefits. In particular, in the field of ML for medical imaging, it allows studying the generalizability of a method. It is thus a step towards its applicability to the clinic. To that aim, the use of multiple datasets is of paramount importance. This will not only allow ruling out that a method is overly specific to a given dataset. It will allow defining which are the bounds within which the method applies. This includes the machine model, the acquisition parameters, and the data quality. It also includes factors which are unrelated to imaging such as population age, sex, geographic origin, disease severity, and others.

6 Measurement Reproducibility

What Is It? Measurement reproducibility is the study of the variability of a specific measurement under different acquisition conditions. We are aware that, at first sight, this concept does not fit ideally in our taxonomy (*see* Subheading 7.1 for a more detailed discussion). Nevertheless, we chose to present it as a separate entity because this is a very common meaning of the word reproducibility in medical imaging[18] (e.g., [62–69]) and we thus believe that it deserves a special treatment. Here, we consider an algorithm that produces a measurement for each individual patient (for instance, the volume of an anatomical structure computed by a segmentation method). A prototypical example of measurement reproducibility is the test-retest reproducibility: how much does the measure vary when applied to two different scans of the same patient? One can then introduce different variations: scans on the same day or not, scans on the same or different machines, systematic addition of noise or artifacts to the data.... Finally, some authors call inter-method reproducibility the comparison of different software packages for the measurement of the same anatomical entity [70]. We do not believe this falls within the topic of reproducibility but rather of methods' comparison.[19]

What Does It Require? The code is necessary to make sure that variations do not depend on implementation and to ease the reproducibility study. The trained models are also very welcome to facilitate the process. It is then necessary to have access to test-retest data, meaning different acquisitions of the same patient. As mentioned above: the more varied these different acquisitions, the more extensive the study. Ideally, one would want to have access to scans performed on the same day [62, 67], on different days [65, 66], at different times during the day (e.g., before or after caffeine consumption, a factor which affects functional MRI measures [71]), on different imaging devices [63], and with different acquisition parameters [68]...It is unlikely to obtain that many scans for the same patients. A more feasible approach is to study these different factors of variations for different patients. Furthermore, starting with a given image, it is possible to simulate different types of alterations and defects by adding them to the original image. This can be very useful because it allows generating very large numbers of images easily and to control for specific imaging characteristics (such as, e.g., the level of noise or the strength of motion artifacts). Such simulations can involve completely synthetic images called phantoms [72] which mimic real images. It can also be done through the addition of defects to real images [73–75]. Ideally, measurement reproducibility should be performed in different populations of participants separately (for instance, a child with autism spectrum disorder or a patient with Parkinson's disease is more likely to move during the acquisition, and the image is thus more likely to be affected by motion artifacts).

[18] Note that the word is used to evaluate reproducibility of automatic methods across different scans of the same subject but also when a human rater is involved (manual or semi-automated measurements), including intra-rater (measurement twice by the same rater) and inter-rater (two different raters) reproducibility from a single scan.

[19] It is of interest to compare which of them is the most accurate or robust, with respect to a ground truth. However, as mentioned above, we do not believe it falls within the topic of reproducibility.

Why Is It Useful? Measurement reproducibility is central for measurement sciences, and medical imaging is one of those. It is an extremely precious information to the user (for instance, the radiologist). Indeed, it provides, at the individual patient level, and ideally for different categories of patients, the precision that they may expect from the measurement tool. There is a wide tradition to perform such reproducibility studies in radiology journals. We believe that it would be very welcome that it becomes more commonplace in the ML for medical imaging community.

7 Discussion

7.1 About the Different Types of Reproducibility

We have presented different types of reproducibility. Our taxonomy is not original nor aims at being universal. The boundaries between types are partly fuzzy. For instance, to which degree replication with a different but similar dataset should be considered *statistical* or *conceptual* reproducibility? We do not believe such questions to be of great importance. Rather, it is fruitful, following Gundersen and Kjensmo [14] and Peng [76], to consider reproducibility as a spectrum. In particular, one can consider that the first three types provide increasing support for a finding: conceptual provides more support than statistical which in turns provides more support than exact. The amount of components necessary to perform them is in the reverse order: exact requires more than statistical which requires more than conceptual. Does it mean that only conceptual reproducibility matters? Absolutely not. As we mentioned, other types of reproducibility are necessary to dissect why a given replication has failed as well as to better specify the bounds within which a scientific claim is valid. Last but not least, exact reproducibility also helps build trust in science.

We must admit that measurement reproducibility does not fit very well in this landscape. Moreover, one could also argue that it is a type of conceptual reproducibility, which is partly true as it aims at studying the reproducibility when varying the input data. We nevertheless believe it deserves a special treatment, for several reasons. First, here reproducibility is studied at the individual (i.e., patient) level and not at the population level. Also, the emphasis is on the measurement rather than the finding. Even if it has its role in the building of scientific knowledge, it has specific practical implications for the user. Moreover, as mentioned above, this is actually the most widely used meaning of reproducibility in medical imaging, and it seemed important that the reader is acquainted with it.

7.2 The Many Benefits of Reproducibility

"Der Weg ist das Ziel" is a German saying which can be roughly translated as: "the path is the goal." Indeed, reproducibility allows researchers to discover many new places down the road before reaching the final destination. Even if this destination is never reached, the benefits of the travel are of major importance. Let us try to list some of them.

There are many individual benefits for researchers and labs. An important one is that aiming at reproducible research results in *reusable* research artifacts. How agreeable it is for a researcher to easily reuse an old code for a new project! How useful it is for a research lab to have data organized according to community standards making it easier to reuse and share! Moreover, papers that come with shared data [22, 77, 78] or code attract [79], on average, more citations. Thus aiming at reproducibility is also in the researchers' self-interest.

There are also considerable benefits for the scientific community as a whole. As mentioned before, reproducible research is often associated to open code, open data, and available trained models. This allows researchers not only to use them to perform replication studies but also to use these research artifacts for completely different purposes such as building new methods or conducting analysis on pooled datasets. In the specific case of ML for medical imaging, it also allows assessing independently the influence of preprocessing, feature extraction, and ML method. This is particularly important when claims of superiority of a new ML method are made, but the original paper uses overly specific preprocessing steps.

Of course, at the end of the path, the goal itself brings many benefits. These have already largely described in the previous sections so we will just mention them briefly. Conceptual replication studies are necessary for corroborating findings and thus building new scientific knowledge. Statistical replication allows ensuring that results are not due to cherry picking. Exact replication allows detecting errors and increases trust in science in general.

7.3 Awareness Is Rising

Throughout this chapter, we have referred to numerous papers, resources, and tools that demonstrate that awareness regarding reproducibility has strongly risen in the past years.

Various papers and studies have highlighted the lack of reproducibility in different fields (e.g., [11, 15, 19]). In machine learning for medical imaging, Simko et al. [21] have studied the reproducibility of methods (mainly code availability and usability thus restricted to exact reproducibility) published at the Medical Imaging with Deep Learning (MIDL) conference from 2018 to 2022 and found that about 20% of papers came with a repository that was deemed reproducible.

Various papers have been published providing advice and guidelines [10, 17, 22–27]. Some of the guidelines include reproducibility checklists. Some checklists are associated to a specific journal or conference and are provided to the reviewers so that they can take these aspects in consideration when evaluating papers. One can cite, for example, the MICCAI (Medical Image Computing and Computer-Assisted Intervention conference) reproducibility checklist.[20,21]

Finally, it is very important that reproducibility studies, assessing all aspects of reproducibility (exact, statistical, conceptual, measurement), are performed, published, and widely read. Unfortunately, it is still easier to publish in a high-impact journal a study that is not reproducible but describes exciting results than a replication study. The good news is that this is starting to change. Reproducibility challenges have been proposed in various fields including machine learning[22] and medical image computing [80]. In the field of neuroimaging, the journal NeuroImage: Reports publishes Open Data Replication Reports.[23] the Organization for Human Brain Mapping has a replication award.[24] and the MRITogether workshop[25] emphasizes reproducibility.

7.4 One Size Does Not Fit All

We hope the reader is now convinced of the benefits of aiming towards reproducible research. Does it mean that reproducibility requirements should be the same for all studies? We strongly believe the opposite. To take an extreme example, requiring all studies to be exactly reproducible with minimal efforts (like with running a single command) would be an awful idea. We believe, on the contrary, that reproducibility efforts should vary according to many factors including the type of study and the context in which it is performed. One would probably not have the same level of requirement for a methodological paper and for an extensive medical application with strong claims about clinical applicability. For the former, one may be satisfied with an experiment on a single or a few datasets. For the later, one would expect the study to include

[20] https://miccai2021.org/files/downloads/MICCAI2021-Reproducibility-Checklist.pdf.

[21] https://github.com/JunMa11/MICCAI-Reproducibility-Checklist.

[22] https://paperswithcode.com/rc2022.

[23] https://www.journals.elsevier.com/neuroimage-reports/infographics/neuroimage-reports-presents-open-data-replication-reports?utm_campaign=STMJ_176479_SC&utm_medium=email&utm_acid=268008024&SIS_ID=&dgcid=STMJ_176479_SC&CMX_ID=&utm_in=DM292849&utm_source=AC_.

[24] https://www.humanbrainmapping.org/i4a/pages/index.cfm?pageid=3731.

[25] https://mritogether.esmrmb.org/.

multiple datasets with varying characteristics and a comprehensive assessment of generalizability under different factors such as imaging devices and acquisition parameters. Also, there are some cases where sharing the code is not desired (e.g., because an industrial application is foreseen) or where the code will not adhere to best development practices because it is just a prototype to test a new methodology. Nevertheless, sharing weakly documented code is always better than no sharing at all. Similarly, there are cases where data sharing is difficult or even impossible due to regulatory constraints. As mentioned above, reproducibility is a spectrum. Where a given study should lie in this spectrum should depend on the type of study and the constraints the researchers face.

We thus advocate for a nondogmatic approach to reproducibility. Guidelines are extremely useful, but they should not be carved in stone. Also, we believe that the requirements should be assessed by the reviewers on a case-by-case basis. Indeed, what matters is that the reproducibility level matches the claims made in the paper. Of course, it is a good thing that journals and conferences provide requirements for reporting essential information. It is helpful to researchers and makes the community progress towards better science. Also, some bad practices such as data leakage or p-hacking need to be banished. But we believe that very high reproducibility requirements (e.g., requiring that exact reproducibility is feasible) at the level of a given journal or conference would be counterproductive. Finally, we like the idea of a badging system [27] which would tag papers according to their reproducibility level. It remains to be seen how such system should be implemented.

To conclude, we firmly believe that it is essential for researchers and students in the field of ML for medical imaging to be trained to the concepts and practice of reproducibility. It will be beneficial to them as well as to the community in general. But this does not mean that researchers should aim at perfect reproducibility in all their studies. Diversity in research approaches and practices is also a factor that drives science forward and which should be preserved.

Acknowledgements

The authors are grateful to G. Varoquaux for pointing them towards useful references. This work was supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA

Institut Hospitalo-Universitaire-6). ETS acknowledges funding from the 4TU Precision Medicine program supported by High Tech for a Sustainable Future, a framework commissioned by the four Universities of Technology of the Netherlands.

References

1. Seab J, Jagust W, Wong S, Roos M, Reed BR, Budinger T (1988) Quantitative NMR measurements of hippocampal atrophy in Alzheimer's disease. Magn Reson Med 8(2):200–208
2. Lehericy S, Baulac M, Chiras J, Pierot L, Martin N, Pillon B, Deweer B, Dubois B, Marsault C (1994) Amygdalohippocampal MR volume measurements in the early stages of Alzheimer disease. Am J Neuroradiol 15(5): 929–937
3. Jack CR, Petersen RC, Xu YC, Waring SC, O'Brien PC, Tangalos EG, Smith GE, Ivnik RJ, Kokmen E (1997) Medial temporal atrophy on MRI in normal aging and very mild alzheimer's disease. Neurology 49(3):786–794
4. Varoquaux G, Colliot O (2022) Evaluating machine learning models and their diagnostic value. HAL preprint hal-03682454. https://hal.archives-ouvertes.fr/hal-03682454/
5. Thibeau-Sutre E, Diaz M, Hassanaly R, Routier A, Dormont D, Colliot O, Burgos N (2022) ClinicaDL: an open-source deep learning software for reproducible neuroimaging processing. Comput Methods Prog Biomed 220:106818
6. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22: 1359–1366
7. Baker M (2016) 1,500 scientists lift the lid on reproducibility. Nature 533:452–454
8. Gundersen OE (2020) The reproducibility crisis is real. AI Mag 41(3):103–106
9. Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2(8): e124
10. Begley CG, Ioannidis JP (2015) Reproducibility in science: improving the standard for basic and preclinical research. Circ Res 116(1): 116–126
11. Collaboration OS (2015) Estimating the reproducibility of psychological science. Science 349(6251):aac4716
12. Begley CG (2013) An unappreciated challenge to oncology drug discovery: pitfalls in preclinical research. Am Soc Clin Oncol Educ Book 33(1):466–468
13. Sonnenburg S, Braun ML, Ong CS, Bengio S, Bottou L, Holmes G, LeCunn Y, Muller KR, Pereira F, Rasmussen CE et al (2007) The need for open source software in machine learning. J Mach Learn Res 8(81):2443–2466
14. Gundersen OE, Kjensmo S (2018) State of the art: reproducibility in artificial intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32
15. Hutson M (2018) Artificial intelligence faces reproducibility crisis. Science 359(6377): 725–726
16. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Waldron L, Wang B, McIntosh C, Goldenberg A, Kundaje A, Greene CS, et al (2020) Transparency and reproducibility in artificial intelligence. Nature 586(7829):E14–E16
17. Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d'Alché Buc F, Fox E, Larochelle H (2021) Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. J Mach Learn Res 22:1–20
18. Stupple A, Singerman D, Celi LA (2019) The reproducibility crisis in the age of digital medicine. NPJ Digit Med 2(1):1–3
19. McDermott M, Wang S, Marinsek N, Ranganath R, Ghassemi M, Foschini L (2019) Reproducibility in machine learning for health. arXiv preprint arXiv:190701463
20. Beam AL, Manrai AK, Ghassemi M (2020) Challenges to the reproducibility of machine learning models in health care. JAMA 323(4): 305–306
21. Simko A, Garpebring A, Jonsson J, Nyholm T, Löfstedt T (2022) Reproducibility of the methods in medical imaging with deep learning. arXiv preprint arXiv:221011146
22. Gorgolewski KJ, Poldrack RA (2016) A practical guide for improving transparency and reproducibility in neuroimaging research. PLoS Biol 14(7):e1002506
23. Nichols TE, Das S, Eickhoff SB, Evans AC, Glatard T, Hanke M, Kriegeskorte N, Milham MP, Poldrack RA, Poline JB et al (2017) Best practices in data analysis and sharing in neuroimaging using MRI. Nat Neurosci 20(3): 299–303

24. Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Nichols TE, Poline JB, Vul E, Yarkoni T (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci 18(2):115–126
25. Niso G, Botvinik-Nezer R, Appelhoff S, De La Vega A, Esteban O, Etzel JA, Finc K, Ganz M, Gau R, Halchenko YO et al (2022) Open and reproducible neuroimaging: from study inception to publication. Neuroimage 263:119623
26. Turkyilmaz-van der Velden Y, Dintzner N, Teperek M (2020) Reproducibility starts from you today. Patterns 1(6):100099
27. Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC (2021) Reproducibility standards for machine learning in the life sciences. Nat Methods 18(10):1132–1135
28. Varoquaux G (2018) Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180:68–77
29. Button KS, Ioannidis J, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14(5):365–376
30. Varoquaux G, Cheplygina V (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med 5(1):1–8
31. Bouthillier X, Laurent C, Vincent P (2019) Unreproducible research is reproducible. In: International Conference on Machine Learning, PMLR, pp 725–734
32. Langer SG, Shih G, Nagy P, Landman BA (2018) Collaborative and reproducible research: goals, challenges, and strategies. J Digit Imaging 31(3):275–282
33. Goodman SN, Fanelli D, Ioannidis JP (2016) What does research reproducibility mean? Sci Transl Med 8(341):341ps12–341ps12
34. Plesser HE (2018) Reproducibility vs. replicability: a brief history of a confused terminology. Front Neuroinfor 11:76
35. McDermott MB, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M (2021) Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 13(586):eabb1655
36. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3(1):1–9
37. Gabelica M, Bojčić R, Puljak L (2022) Many researchers were not compliant with their published data sharing statement: mixed-methods study. J Clin Epidemiol 150:33–41
38. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, et al (2016) The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data 3(1):1–9
39. Bourget MH, Kamentsky L, Ghosh SS, Mazzamuto G, Lazari A, Markiewicz CJ, Oostenveld R, Niso G, Halchenko YO, Lipp I, et al (2022) Microscopy-BIDS: an extension to the Brain imaging data structure for microscopy data. Front Neurosci 16: e871228
40. Saborit-Torres J, Saenz-Gamboa J, Montell J, Salinas J, Gómez J, Stefan I, Caparrós M, García-García F, Domenech J, Manjón J, et al (2020) Medical imaging data structure extended to multiple modalities and anatomical regions. arXiv preprint arXiv:201000434
41. Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO, Chupin M, Benali H, Colliot O (2011) Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56(2): 766–781
42. Samper-González J, Burgos N, Bottani S, Fontanella S, Lu P, Marcoux A, Routier A, Guillon J, Bacci M, Wen J, et al (2018) Reproducible evaluation of classification methods in Alzheimer's disease: framework and application to MRI and PET data. Neuroimage 183:504–521
43. Karakuzu A, DuPre E, Tetrel L, Bermudez P, Boudreau M, Chin M, Poline JB, Das S, Bellec P, Stikov N (2022) Neurolibre: a preprint server for full-fledged reproducible neuroscience. OSF Preprints
44. Crane M (2018) Questionable answers in question answering research: reproducibility and variability of published results. Trans Assoc Comput Linguist 6:241–252
45. Jungo A, Scheidegger O, Reyes M, Balsiger F (2021) pymia: a python package for data handling and evaluation in deep learning-based medical image analysis. Comput Methods Program Biomed 198:105796. https://doi.org/10.1016/j.cmpb.2020.105796
46. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U, et al (2021) Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21), pp 2633–2650

47. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp 308–318
48. Gorgolewski KJ, Alfaro-Almagro F, Auer T, Bellec P, Capotă M, Chakravarty MM, Churchill NW, Cohen AL, Craddock RC, Devenyi GA, et al (2017) BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. PLoS Comput Biol 13(3):e1005209
49. Routier A, Burgos N, Díaz M, Bacci M, Bottani S, El-Rifai O, Fontanella S, Gori P, Guillon J, Guyot A, et al (2021) Clinica: an open-source software platform for reproducible clinical neuroscience studies. Front Neuroinform 15:e689675
50. McCormick M, Liu X, Jomier J, Marion C, Ibanez L (2014) ITK: enabling reproducible research and open science. Front Neuroinform 8:13
51. Yoo TS, Ackerman MJ, Lorensen WE, Schroeder W, Chalana V, Aylward S, Metaxas D, Whitaker R (2002) Engineering and algorithm design for an image processing API: a technical report on ITK-the insight toolkit. In: Medicine Meets Virtual Reality 02/10, IOS press, pp 586–592
52. Drummond C (2009) Replicability is not reproducibility: nor is it good science. In: Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML, vol 1
53. Bouthillier X, Delaunay P, Bronzi M, Trofimov A, Nichyporuk B, Szeto J, Mohammadi Sepahvand N, Raff E, Madan K, Voleti V, et al (2021) Accounting for variance in machine learning benchmarks. Proc Mach Learn Syst 3:747–769
54. Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durrleman S, Burgos N, Colliot O (2020) Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med Image Anal 63:101694
55. Samala RK, Chan HP, Hadjiiski L, Koneru S (2020) Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks. In: Proceedings of SPIE Medical Imaging 2020: Computer-Aided Diagnosis, International Society for Optics and Photonics, vol 11314, p 1131416
56. Panwar H, Gupta PK, Siddiqui MK, Morales-Menendez R, Singh V (2020) Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. Chaos, Solitons Fractals 138:109944
57. Bussola N, Marcolini A, Maggio V, Jurman G, Furlanello C (2021) AI slipping on tiles: Data leakage in Digital Pathology. In: Del Bimbo A, Cucchiara R, Sclaroff S, Farinella GM, Mei T, Bertini M, Escalante HJ, Vezzani R (eds) Pattern recognition. ICPR International Workshops and Challenges, Springer International Publishing, Cham. Lecture notes in computer science, pp 167–182. https://doi.org/10.1007/978-3-030-68763-2_13
58. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of p-hacking in science. PLoS Biol 13(3):e1002106
59. Henderson EL (2022) A guide to preregistration and registered reports. Preprint. https://osf.io/preprints/metaarxiv/x7aqr/download
60. Bottani S, Burgos N, Maire A, Wild A, Ströer S, Dormont D, Colliot O, Group AS et al (2022) Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse. Med Image Anal 75:102219
61. Perkuhn M, Stavrinou P, Thiele F, Shakirin G, Mohan M, Garmpis D, Kabbasch C, Borggrefe J (2018) Clinical evaluation of a multiparametric deep learning model for glioblastoma segmentation using heterogeneous magnetic resonance imaging data from clinical routine. Investig Radiol 53(11):647
62. Lukas C, Hahn HK, Bellenberg B, Rexilius J, Schmid G, Schimrigk SK, Przuntek H, Köster O, Peitgen HO (2004) Sensitivity and reproducibility of a new fast 3D segmentation technique for clinical MR-based brain volumetry in multiple sclerosis. Neuroradiology 46(11):906–915
63. Borga M, Ahlgren A, Romu T, Widholm P, Dahlqvist Leinhard O, West J (2020) Reproducibility and repeatability of MRI-based body composition analysis. Magn Reson Med 84(6):3146–3156
64. Chard DT, Parker GJ, Griffin CM, Thompson AJ, Miller DH (2002) The reproducibility and sensitivity of brain tissue volume measurements derived from an SPM-based segmentation methodology. J Magn Reson Imaging: Off J Int Soc Magn Reson Med 15(3):259–267
65. de Boer R, Vrooman HA, Ikram MA, Vernooij MW, Breteler MM, van der Lugt A, Niessen WJ (2010) Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods. Neuroimage 51(3):1047–1056

66. Lemieux L, Hagemann G, Krakow K, Woermann FG (1999) Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. Magn Reson Med: Off J Int Soc Magn Reson Med 42(1):127–135
67. Tudorascu DL, Karim HT, Maronge JM, Alhilali L, Fakhran S, Aizenstein HJ, Muschelli J, Crainiceanu CM (2016) Reproducibility and bias in healthy brain segmentation: comparison of two popular neuroimaging platforms. Front Neurosci 10:503
68. Yamashita R, Perrin T, Chakraborty J, Chou JF, Horvat N, Koszalka MA, Midya A, Gonen M, Allen P, Jarnagin WR et al (2020) Radiomic feature reproducibility in contrast-enhanced CT of the pancreas is affected by variabilities in scan parameters and manual segmentation. Euro Radiol 30(1):195–205
69. Poldrack RA, Whitaker K, Kennedy DN (2019) Introduction to the special issue on reproducibility in neuroimaging. Neuroimage 218:116357
70. Palumbo L, Bosco P, Fantacci M, Ferrari E, Oliva P, Spera G, Retico A (2019) Evaluation of the intra-and inter-method agreement of brain MRI segmentation software packages: a comparison between SPM12 and FreeSurfer v6. 0. Phys Med 64:261–272
71. Laurienti PJ, Field AS, Burdette JH, Maldjian JA, Yen YF, Moody DM (2002) Dietary caffeine consumption modulates fMRI measures. Neuroimage 17(2):751–757
72. Collins DL, Zijdenbos AP, Kollokian V, Sled JG, Kabani NJ, Holmes CJ, Evans AC (1998) Design and construction of a realistic digital brain phantom. IEEE Trans Med Imaging 17(3):463–468
73. Shaw R, Sudre C, Ourselin S, Cardoso MJ (2018) MRI K-space motion artefact augmentation: Model robustness and task-specific uncertainty. In: Medical Imaging with Deep Learning – MIDL 2018
74. Duffy BA, Zhang W, Tang H, Zhao L (2018) Retrospective correction of motion artifact affected structural MRI images using deep learning of simulated motion. In: Medical Imaging with Deep Learning – MIDL 2018
75. Loizillon S, Bottani S, Maire A, Ströer S, Dormont D, Colliot O, Burgos N (2023) Transfer learning from synthetic to routine clinical data for motion artefact detection in brain t1-weighted MRI. In: SPIE Medical Imaging 2023: Image Processing
76. Peng RD (2011) Reproducible research in computational science. Science 334(6060): 1226–1227
77. Piwowar HA, Day RS, Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. PLoS One 2(3): e308
78. Piwowar HA, Vision TJ (2013) Data reuse and the open data citation advantage. PeerJ 1:e175
79. Vandewalle P (2012) Code sharing is associated with research impact in image processing. Comput Sci Eng 14(4):42–47
80. Balsiger F, Jungo A, Chen J, Ezhov I, Liu S, Ma J, Paetzold JC, Sekuboyina A, Shit S, Suter Y et al (2021) The miccai hackathon on reproducibility, diversity, and selection of papers at the miccai conference. arXiv preprint arXiv:210305437