Reproducibility in medical image computing: what is it and how is it assessed?

Olivier Colliot, Elina Thibeau-Sutre, Camille Brianceau, Ninon Burgos

Chapter 9

Reproducibility in medical image computing: what is it and how is it assessed?

Olivier Colliot[a,*], **Elina Thibeau-Sutre**[b], **Camille Brianceau**[a], **and Ninon Burgos**[a,*]

[a]Sorbonne Université, Institut du Cerveau – Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, F-75013, Paris, France
[b]Department of Applied Mathematics, Technical Medical Centre, University of Twente, Enschede, The Netherlands
[*]Correspondence to: olivier.colliot@cnrs.fr and ninon.burgos@cnrs.fr

Abstract

Medical image computing (MIC) is devoted to computational methods for analysis of medical imaging data and their assessment through experiments. It is thus an experimental science. Reproducibility is a cornerstone of progress in all experimental sciences. As in many other fields, there are major concerns that reproducibility is unsatisfactory in MIC. However, reproducibility is not a single concept but a spectrum, which is often misunderstood by researchers. Moreover, even though some measures have been put in place to promote reproducibility in the MIC community, it is unclear if they have been effective so far.

The objectives of the present chapter are three-fold: i) to provide readers with the necessary concepts underlying reproducibility in MIC; ii) to describe the measures which have been put in place and assess some of them; iii) to sketch some possible new actions that could be taken.

The chapter starts with the presentation of a conceptual framework which distinguishes between different types of reproducibility as well as of the main building blocks of reproducible research. Then, a description of how reproducibility is currently assessed at the MICCAI (Medical Image Computing

In, M. Lorenzi and M. Zuluaga (Eds.), *Trustworthy AI for medical imaging*, Elsevier
https://XXXX

and Computer Assisted Interventions) conference is presented. In particular, we perform a quantitative analysis of MICCAI reviews. It reveals that, on the matter of reproducibility, reviews are unreliable and uninformative. Furthermore, some bad practices of some of the authors are unveiled. Finally, we summarize the current state of affairs and suggest some potential actions that could be discussed within the community to progress towards more reproducible research. In particular, it is very important to have in mind that reproducibility is a spectrum, that there will never be a "one-size-fits-all" model but that there is plenty of room for improvement across all types of reproducibility.

The code and data to reproduce the results of this paper are available at: github.com/reproducibility-reviews/reproducibility-reviews.

Keywords: reproducibility, replicability, reliability, medical imaging, machine learning, deep learning, artificial intelligence

1. Introduction

Medical image computing (MIC) is the field devoted to computational methods for the analysis of medical imaging data. As such, it comprises two main components: i) the design of new methodologies; ii) experimental work to assess the value of these methodologies. It is thus an experimental science, even though many of its experiments are digital.

Reproducibility is at the core of the scientific method in experimental sciences. Indeed, in itself, a scientific study produces findings which are expressed as claims in a paper. For such findings to become knowledge, they need to be reproduced in other studies. Let's take for example the following finding: "this neural network [as described in a given paper] using MRI data as input is able to distinguish patients with Alzheimer's disease from healthy controls with an accuracy superior to 80%". For this finding to become knowledge, it needs to be subsequently reproduced in other studies. When this has been done several times, one can add this to the existing body of knowledge and science can continue to advance on solid ground. Of course, such a finding will usually not be universal: does it hold for different age ranges? for different MRI sequence parameters? etc. Reproducing the study thus also allows clarifying the boundaries within which the knowledge holds.

One can see that reproducibility is crucial in MIC as in any experimental science. However, one specificity of MIC needs to be mentioned: its ultimate goal is clinical translation to improve patient care. There is wide consensus that too little MIC research ultimately leads to clinical advances [1]. Reproducibility issues may be one of the underlying causes [2].

Numerous scientists are concerned about lack of reproducibility of studies [3]. Such concern is obviously not specific to MIC and is present in many fields [4, 5, 6, 7, 8]. It is interesting to note that the concern has grown in the field of machine learning (ML) [9, 10, 11, 12, 13, 14] since most modern MIC methods are based on ML. Nevertheless, MIC has specificities that may differ from general ML: (relatively) small sample size, need to clinically characterize studied samples, variability in image sequences and devices, differences between research data and clinical routine data, difficulty of ground-truth generation... Several papers have thus been published on the specific issue of reproducibility in MIC [15, 16, 17] or in closely related fields such as computational pathology [2].

Concerns about reproducibility encompass different aspects. In particular, one can distinguish: i) failures to reproduce previous papers; ii) the fact that many papers do not provide sufficient information for reproduction. The two may be related but there is no bijection: a paper may include all the necessary information and reproduction attempts fail (for instance, because the original study had too many degrees of freedom and at the end "overfitted" a given dataset).

More generally, reproducibility is not a single concept but a spectrum: "reproducing a study" may mean different things including but not limited to: i) obtaining exactly the same results as in the original paper; ii) obtaining similar results using a different dataset; iii) studying the generalization under variations of testing data and/or characteristics of the method [13]. Moreover, different words have been used to describe reproducibility including replicability, repeatability, reliability and others. There is thus a need for a clear conceptual framework distinguishing between different types of reproducibility. Such a framework is proposed in Section 2.

Reproducibility requires different components either within the paper itself or items that may come together with the paper (code, data, etc). This entails including all necessary information in the paper but also providing access to data, code or other elements. Specifically, these *building blocks of reproducibility* are described in Section 3. Assessing these different blocks allows assessing the reproducibility of a study.

Assessment of reproducibility has been included as a criterion in the review process for submissions to the MICCAI (Medical Image Computing and Computer-Assisted Intervention) conference. Namely, submissions are required to come with a reproducibility checklist and reproducibility is an item that needs to be assessed by the reviewers. These are very valuable initiatives and reflects that awareness of reproducibility issues has risen in the field of MIC (and in the related field of computer-aided interventions [CAI] as well). However, it seems of interest to now take a look at the review process and see if it provides adequate information to assess reproducibility of papers. We thus

conducted an original analysis of the MICCAI 2023 reviews on reproducibility. The corresponding results are presented in Section 4.

To conclude, we summarize what is the state of affairs regarding reproducibility in MIC (Section 5) and suggest some possible actions for improving both the reproducibility of research in our field and its assessment within conferences and journals (Section 6). Obviously, these are only suggestions aiming to stimulate discussion within the community.

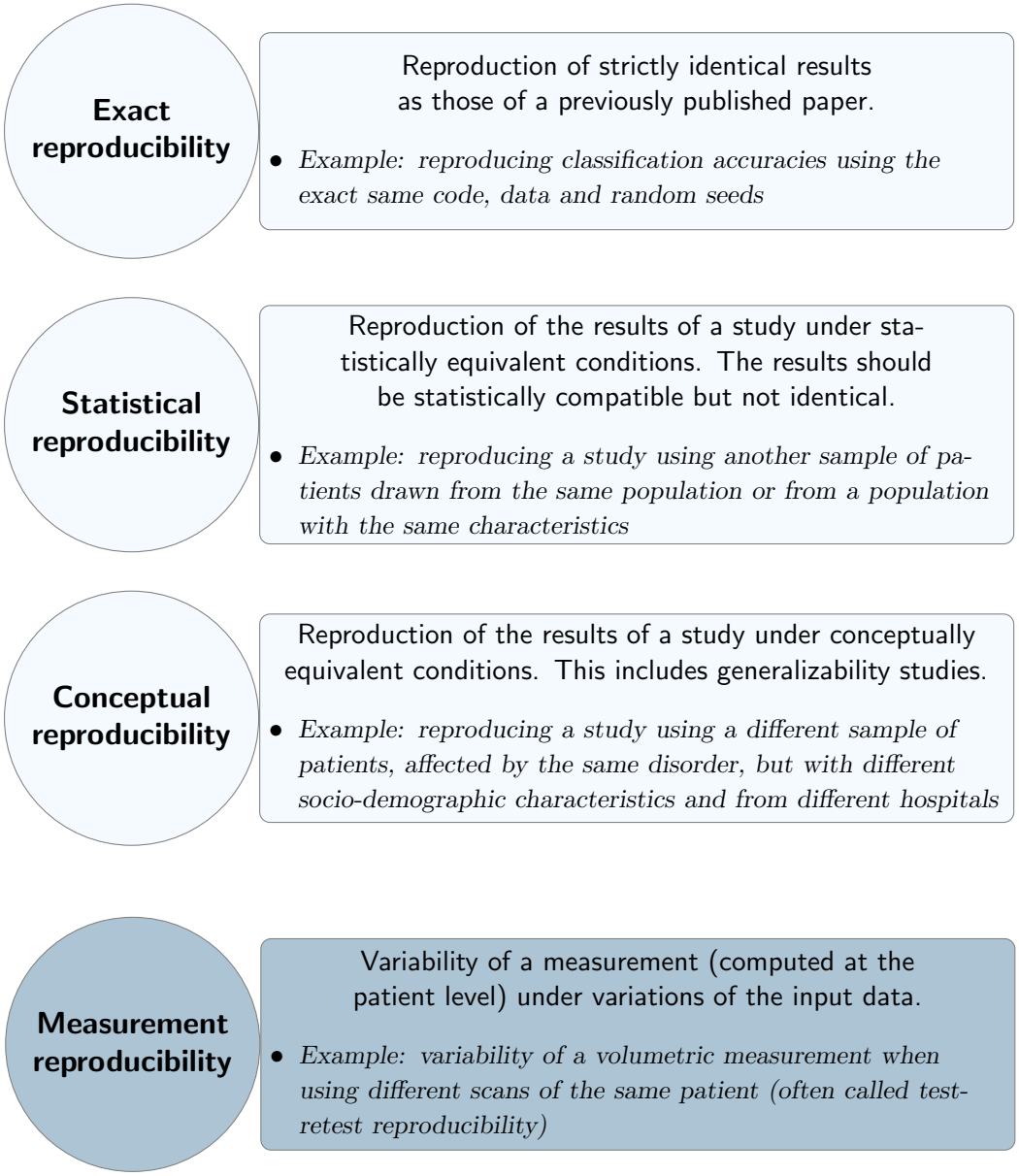

Figure 1: Different types of reproducibility. Adapted from [17] (CC BY 4.0).

Box 1: Glossary

- **Reproducibility, replicability, repeatability.** These nouns will be used as synonyms of reproducibility. Similarly, *to replicate* will be used as a synonym of *to reproduce*.
- **Original study.** Study that first showed a finding.
- **Replication study.** Study that subsequently aimed at replicating an original study.
- **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.
- **Limitations.** A set of restrictions under which the claims may not hold (usually because the corresponding settings have not been explored).
- **Research artifact.** Any output of scientific research: papers, code, data, protocols...
- **Method.** The 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 openneuro.org).
- **Semi-public data.** Data which requires approval of a research project (for instance the Alzheimer's Disease Neuroimaging Initiative [ADNI] www.adni-info.org). The researchers can then use the data only for the intended research purpose and cannot redistribute it.
- **Trained models**. Machine learning (ML) models trained in the original study.
- **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 [18, 19] 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 [20]. 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.

Adapted from [17] (CC BY 4.0).

2. Conceptual framework

This section presents a framework which distinguishes between different types of reproducibility, thereby providing a *taxonomy* of reproducibility. It has been presented in a previous work [17] which itself took inspiration from existing papers [13, 21, 9, 22, 23][1].

We propose to distinguish between four types of reproducibility: **exact reproducibility**, **statistical reproducibility**, **conceptual reproducibility** and **measurement reproducibility**. The first three types are related: they all concern the ability to reproduce the findings of a previous study. The fourth is different as it concerns the reproducibility of a *measurement* rather than that of a *study*. They are summarized in Figure 1.

One should note that, in the literature, various terms have been used to refer to reproducibility [23]: replicability, repeatability, reliability, robustness, generalizability... Some authors have proposed to use specific words to refer to different types of reproducibility but there is no consensus. Even worse, some of these words, for instance reproducibility vs replicability, have even been used by some authors with opposite meanings as reported in [23, 24]. For sake of clarity, we advocate to systematically use the term reproducibility, with an adequate adjective to specify what type of reproducibility one refers to. As argued in [23], we believe it is more efficient than trying to assign different meanings to words which are quasi-synonyms in common language.

The remainder of this section presents the different types of reproducibility, in a concise manner. For a more extensive description, the reader is referred to our previous publication [17]. Some of the terms used are defined in the glossary in Box 1.

2.1 Exact reproducibility

Exact reproducibility aims at reproducing the exact results reported in a published paper, following the same procedures. If things go well, one will obtain the same figures and tables.

It requires to have access to data, code, experimental procedures (e.g. data splits, criteria for model selection). Note that when there are non-deterministic components, random seeds need to be stored. It also makes things easier if trained models are shared since it allows comparing the results of models trained by the authors to those of models retrained during replication.

Exact reproducibility has many benefits. First, it allows detecting fraud. It is difficult to know the prevalence of fraud. One can only hope that outright

[1] In particular, we got inspired by the idea proposed in [23] to add adjectives to distinguish between types of reproducibility and three of the four types that we distinguish are close to those proposed in [13, 21], even though not identical.

fraud is rare. But, in any case, it has disastrous consequences. Second, while one can hope that fraud is rare, we all make mistakes. Exact reproducibility eases the detection of errors: it is a service to the community and to the authors themselves. It can help discover insidious errors such as "biases and artifacts in the data that were missed by the authors and that cannot be discovered if the data are never made available" [25], subtle data leakage or implementation errors that make the code inconsistent with the paper. It should not aim at pointing fingers. On the contrary, it should help us all progress towards a more transparent, trustworthy and peaceful way of doing research. Finally, a positive side effect of exact reproducibility is that it comes with *open science artifacts* (e.g. data, code) which has many benefits beyond reproducibility.

However, one needs to acknowledge that exact reproducibility is not achievable, and not even desirable, for all MIC studies, if only because some medical data cannot be shared easily. We should not aim for "one-size-fits-all" research practices but tailor the reproducibility requirements to the type of research study.

2.2 Statistical reproducibility

Statistical reproducibility aims at reproducing findings of a study under "statistically equivalent" conditions. The definition is less precise than that of exact reproducibility. Nevertheless, the following requirements appear reasonable. The data should be statistically equivalent to the one from the original study. For instance, another subsample of a larger source population or another dataset which characteristics (e.g. age, clinical status, scanner types) are similar to the original ones. Random components should be left at random: this includes for instance random seeds, initialization, data splits, and data order when training. The method and the implementation should be the same as that of the original study, even though one could consider that studying robustness to variations of hyperparameters or minor architectural choices would fall under the umbrella of statistical reproducibility.

It is essential that original studies report error margins on their estimates (through confidence intervals or standard-error). Indeed, one needs to assess if the results of the replication are *statistically compatible* with those of the original study: typically checking whether confidence intervals overlap. As for exact reproducibility, code needs to be accessible to ensure that variations do not come from reimplementation. Data, trained models, random components are not required but would be welcome to further dissect why a replication attempt may have failed.

Statistical reproducibility is essential because MIC is an experimental science. It allows studying robustness to test data, to training splits and to random components. These would even be welcome in the original paper.

Moreover, it is of high interest to attempt replication using a different dataset with statistically equivalent characteristics. Unsuccessful replication may be an indication that the original study overfitted their dataset through excessive experimentation (e.g. trying with different architectures or hyper-parameters). This would indicate that there were too many *researcher degrees of freedom* [20, 26], a concept that will be described in Section 3.

2.3 Conceptual reproducibility

Conceptual reproducibility aims at validating the findings under conceptually similar conditions. This means that the method, the data and the experiments are *compatible* with the claims of the original study but they are not identical.

In principle, one only needs the original paper. Nevertheless, this assumes that reporting has followed best practices including clear description of the methods, the datasets and the experimental procedures. In practice, access to code will considerably ease replication attempts. Furthermore, availability of all components necessary for exact reproducibility (e.g. data, random seeds, preprocessing) allows dissecting replication failures and is thus very helpful.

Conceptual reproducibility allows verifying the claims of the original study but also studying its limitations and ultimately clarifying the boundaries within which the findings hold. It is thus an essential step towards the consolidation of scientific knowledge. It is also a step towards translation to the clinic.

2.4 Measurement reproducibility

Measurement reproducibility aims at studying the robustness of a given measurement when acquisition conditions vary. This concept is quite separated from the first three. However, it corresponds to a common meaning of *reproducibility* in medical imaging. Typically, measurement reproducibility is relevant to MIC methods that result in a measure for each individual patient: for instance a segmentation algorithm that outputs the volume of a lesion or structure. A common case is test-retest reproducibility: one studies how the measure varies for different acquisitions of the same patient.

It requires access to the code and ideally to the trained models since the focus here is on reproducing the output measure, not the training. One needs to have different acquisitions of the same patient that will constitute the test-retest data. The study will be more extensive if acquisition conditions are varied: performed on the same day [27], on different days [28], at different times during the day [29], on different scanners [30], with different acquisition parameters [31]... In the absence of such variations, one can simulate alterations and defects [32, 33, 34]. This cannot completely replace real data but can be useful and has the additional advantage of being able to precisely control the amount of artifacts introduced.

There is a long tradition of studying measurement reproducibility in medical imaging. It is obviously essential for MIC which is often a measurement science. Again, it is a step towards clinical translation as it informs the clinician about the reliability they may expect from the measurement. Measurement reproducibility is often studied in papers published in radiology journals. It definitely also has its place in MIC publications.

3. The building blocks of reproducibility

To attempt replication for each type of reproducibility described above, different components are needed. These building blocks of reproducibility are described below. This takes inspiration from reproducibility checklists [12, 16][2] as well as from previous papers [26, 22, 35, 15, 13, 25, 36, 37].

3.1 Paper

A paper with all the necessary information is required for replication. This sounds trivial. And yet, this is not always the case (by far!). The **method** needs to be precisely and completely described, either within the text or using adequate references. The **data** need to be thoroughly documented. A critical point is the medical description of the data (including but not limited to socio-demographic and clinical characteristics of the participants, image acquisition settings, quality control): doing MIC should not simply be playing with medical images and computer vision tools. Unfortunately, this is often overlooked: many MIC papers do not even report demographic information [1]. **Experimental results** are also often not reported properly. Metrics should be unambiguous: their name should follow consensus guidelines [38] or, if a given metric is not defined in the guidelines, its exact equation needs to be provided. Descriptive and inferential statistics need to be clearly described. Unfortunately, they are often lacking and when a result is presented, it is common to only find a $\pm$ sign, leaving the reader to wonder whether this is a standard-deviation or a standard-error and how it was computed. Similarly, statistical tests need to be precisely specified.

3.2 Code

Code availability is of course very important [13, 26, 25]. Sharing code is nice, sharing documented code is better [14]: at least include a `README` describing how to run it. In the same spirit, it is necessary to specify all dependencies (and versions) and to make their installation as easy as possible [25] for instance

[2] miccai2021.org/files/downloads/MICCAI2021-Reproducibility-Checklist.pdf

using pip [3]. Following good coding practices, including the use of a versioning system or even continuous integration [14], will lead to more robust code and easier maintenance, even though it is clear that not all this can be achieved for a prototype code coming with a new paper. It is a nice idea to include code that allows generating the exact tables and figures included in the paper (e.g. as `.tex` and `.pdf`/`.png` files). One part of the code that is often missing is preprocessing. It is critical to include it as the results depend on it. We refer the reader to *Tips for Publishing Research Code*[4] for further general advice on how to release code accompanying a research paper.

3.3 Data

Access to data is of course necessary to achieve *exact reproducibility*. Whenever possible, it is an excellent initiative to release the data, as it not only serves reproducibility but can be used for other scientific advances. However, being realistic, not all medical data can be shared and data sharing is often in the hands of the data provider rather than in those of MIC researchers. It is important to distinguish between research data, which are acquired as part of a research protocol, and routine clinical data, which are acquired as part of the routine clinical care of the patients. The former can often be made *public* or *semi-public*[5] provided adequate participant consent and ethics approval. However, again, this is often not within the power of MIC researchers. They can nevertheless lobby upon their clinical collaborators to push towards more open data.

A large part of MIC research reuses *public* or *semi-public* data. In that case, it is necessary to specify which participants and which scans have been included. This can be done by providing scan IDs or, more conveniently, by providing code to automatically make the data selection [39]. However, note that, for some datasets such as the UKBiobank[6], the participant IDs are randomly generated for each data request and thus cannot be shared.

A key component to ease reproducibility is to adhere to community standards for data organization. There is a standard for brain imaging, known as BIDS (Brain Imaging Data Structure) [40][7]. More general standards are nevertheless necessary for the MIC community, even though some preliminary steps have been taken with a proposal to extend BIDS to other organs (MIDS - Medical Imaging Data Structure [41][8]).

[3] pypi.org/project/pip
[4] github.com/paperswithcode/releasing-research-code
[5] See glossary Box 1.
[6] www.ukbiobank.ac.uk
[7] bids.neuroimaging.io
[8] See BIDS extension proposal (BEP) number 25 (BEP025) bids.neuroimaging.io/get_involved.html#extending-the-bids-specification

3.4 Training procedure

Most of current MIC relies on ML. It is thus essential to provide all aspects necessary for training models. This should be done both in the paper and in the code. In particular, one needs to specify the data splits, the criteria for model selection, the hyperparameter search procedure and the optimal hyperparameters selected. Moreover, exact reproducibility requires to store random seeds to reproduce non-deterministic aspects [19][9]. Finally, it is good practice to specify which CPUs/GPUs were used and what was the runtime.

3.5 Trained models

It is also a good idea to share trained models. If all the above elements are provided, they could in principle be retrained. Trained models are nevertheless very useful for different aspects of reproducibility: to dissect failures of an *exact reproducibility* attempt, to perform *statistical, conceptual or measurement* reproducibility.

3.6 Statistical analysis

Statistical analysis is very often overlooked in MIC papers. Ideally, one should provide both descriptive and inferential statistics. Descriptive statistics (e.g. standard-deviation, inter-quartile range, box-plots, violin-plots) allows displaying the variability of the performance across a testing set or across training runs. Of note, plots always need to come with a clear caption (e.g. what the box and the whiskers display). Inferential statistics are essential to characterize the precision of the estimates and are necessary for statistical reproducibility. These include confidence interval or standard-error associated to an estimate. One can also perform statistical testing to compare models or methods but this needs to be done with care as many standard statistical tests become invalid in the context of cross-validation [42, 43, 44]. A simpler approach is to statistically compare models on an independent test set because this allows using classical statistical tests (while of course still checking that their assumptions hold).

3.7 Researchers' degrees of freedom

When a researcher has performed excessive experimentation (e.g. many architectures, many hyper-parameters), the likelihood of a successful *statistical or conceptual replication* decreases. These different experiments are referred to as the *researcher degrees of freedom* [20, 26]. This concept is related (but more insidious and thus possibly less faulty) to that of *p-hacking* [45] which consists

[9] clinicadl.readthedocs.io

in performing many statistical tests until something statistically significant pops up. Preventing excessive *researcher degrees of freedom* is difficult. At the very least, one needs to have an independent test set that is separated from the beginning and used only to evaluate the final model. Nevertheless, this is not a bullet-proof solution to *conceptual reproducibility* as one may have overfitted the characteristics of the dataset. Whenever possible, it is nice to have an additional external testing set coming from a different study and thus with different statistical characteristics.

3.8 More realistic datasets

Working with public and benchmark datasets is good because it allows comparing methods in a fair manner and achieving *exact reproducibility*. However, datasets used in MIC papers are very often not representative of clinical routine [1] because they come from research studies with harmonized acquisition protocols and specific inclusion criteria. This reduces the likelihood of *conceptual reproducibility* and ultimately of clinical translation. Thus, more studies based on clinical routine data [46, 47] would be welcome.

3.9 Claims and limitations

Each paper should come with claims and limitations. The claims are one or several statements that describe what can be concluded from the experiments. The claims need to be backed-up by evidence. Typically, one cannot say that "the proposed method outperformed the baseline" by simply reporting point estimates without statistical analysis. It is also essential to mention the limitations, a part that is missing in many MIC papers. The limitations allow delineating the boundaries within which the claims hold (e.g. restricted to research data, to a specific scanner). With clearly stated claims and limitations, one paves the way for well-designed replication studies that can assess how robust the findings are.

4. Assessment of reproducibility at MICCAI

In this section, we aim to characterize the way reproducibility is currently assessed at MICCAI. We took MICCAI as an example because, to our knowledge, no other conference (e.g. ISBI[10], IPMI[11], MIDL[12], SPIE MI[13]) or journal

[10]International Symposium on Biomedical Imaging
[11]Information Processing in Medical Imaging
[12]Medical Imaging With Deep Learning
[13]SPIE Medical Imaging

(e.g. Medical Image Analysis, IEEE TMI[14], JMI[15]) of our community has yet implemented an assessment of reproducibility within its review process.

Section 4.1 first briefly recapitulates the history of reproducibility at MICCAI. Section 4.2 then describes the current reproducibility review process and in particular its reproducibility checklist. The reader is invited to take the place of the reviewer by assessing the reproducibility of "fake" papers that three of the authors created for the *Reproducibility Tutorial*[16] they organized at MICCAI 2023 (Section 4.3). Finally, we conducted an analysis of reproducibility reviews of MICCAI 2023 (Section 4.4).

4.1 The origins: the MICCAI 2020 hackathon

In 2020, the first MICCAI hackathon was held[17]. That year's topic was "reproducibility, diversity, and selection of papers". Their investigations led to the proposition of measures which were divided into immediate and long-term measures [16]. Some of these measures have been fully implemented, some partially, some not at all. The measures and their implementation status are described in Table 1 for immediate measures and in Table 2 for long-term measures.

4.2 MICCAI reproducibility checklist and review instructions

Currently, reproducibility at MICCAI is assessed as follows. One can find the following in the *Paper submission and rebuttal guidelines*[18]:

> "MICCAI is committed to reproducible research. In MICCAI 2023, we strongly encourage authors to improve the reproducibility of their research along three directions: open data, open implementations, and appropriate evaluation design and reporting. Where possible, we invite authors to use open data or to make their data and/or code available for open access by other researchers. Upon submission, authors will be asked to fill out a reproducibility checklist indicating to what extent their submission fulfills these criteria. We encourage reviewers and Area Chairs to take reproducibility of the work into account when assessing a submission."

In the *Reviewer guidelines*[19], one can find the following:

> "Comment on the reproducibility of the paper. Where possible, we encourage authors to use open data or to make their data and code available for open

[14] IEEE Transactions on Medical Imaging
[15] Journal of Medical Imaging
[16] miccai2023-reproducibility-tutorial.github.io
[17] 2020.miccai-hackathon.com
[18] conferences.miccai.org/2023/en/PAPER-SUBMISSION-AND-REBUTTAL-GUIDELINES.html
[19] conferences.miccai.org/2023/en/REVIEWER-GUIDELINES.html

Table 1: Immediate measures proposed by the MICCAI 2020 hackathon.

Measure	Description	Implemented?
Reproducibility checklist	A reproducibility checklist[a] was incorporated at paper submission. The checklist takes inspiration from that of NeurIPS [12] but has been adapted to the specificities of MIC. The authors are asked to fill the checklist at submission. The filled checklist is available to the reviewers. To our knowledge, checklists are not made available together with published papers.	**Yes**
Reproducibility chair	The authors proposed that the reproducibility chair would be "in charge of analyzing the influence of the reproducibility checklist [...], gather feedback from authors and reviewers, and adapt the checklist [...]".	**No**
Statement of data availability	The MICCAI guidelines for authors currently indicate: "Data use declaration and acknowledgment: Authors must declare the data origin, data license, and (when appropriate) ethics application number for any public or private data used in the preparation of the paper."[b] However, to the best of our knowledge, this information is not available once the papers are published.	**Partially**
Promote reproducibility efforts of authors	They proposed to create an official list of open source MICCAI papers. Links to code are now present on the MICCAI website for each paper[c]. However, the visibility is limited since there is no curated list of papers with code. Moreover, the information may be misleading since some of the links are actually broken or lead to empty repositories (see our analysis in Section 4.4.3). Some unofficial lists have been created[d,e]. Furthermore, disseminating code is only one aspect of reproducibility. Links to datasets are also mentioned when available but again this information is not particularly visible. Finally, there is more to reproducibility than only sharing code and datasets. Overall, promotion of reproducibility efforts appears very limited.	**Partially**
Communicate best practices on reproducibility and code submission	They proposed to introduce guidelines for reproducibility. We do not believe this is currently part of the submission guidelines or more generally that there are consensus best practices for reproducibility in the field (except that for validation metrics [38])	**No**

[a] miccai2021.org/files/downloads/MICCAI2021-Reproducibility-Checklist.pdf

[b] conferences.miccai.org/2023/en/PAPER-SUBMISSION-AND-REBUTTAL-GUIDELINES.html (section "3. Manuscript Preparation and Submission")

[c] conferences.miccai.org/2023/papers

[d] github.com/JunMa11/MICCAI-OpenSourcePapers

[e] github.com/yiqings/MICCAI2022_paper_with_code

Table 2: Long-term measures proposed by the MICCAI 2020 hackathon.

Measure	Description	Implemented?
Reproducibility award	A reproducibility award would be part of the promotion of efforts regarding reproducibility.	**No**
Code submission	The authors proposed that code is submitted (in an anonymized way) together with the paper for review. This is currently rarely done even though some guidelines are provided[a].	**No**
Best practices for evaluation	Evaluation comprises several aspects including: i) which metrics are adequate for a given task?; ii) how to appropriately estimate these metrics? Regarding the former, consensus guidelines called "Metrics reloaded" have been published [38]. Regarding the latter, guidelines are lacking.	**Partially**

[a] conferences.miccai.org/2023/en/PAPER-SUBMISSION-FAQ.html (section "Supplementary material")

> access by other researchers. We understand that due to certain restrictions, some researchers are not able to release their proprietary dataset and code; therefore, a clear and detailed description of the algorithm, its parameters, and the dataset is highly valuable. Please provide comments about whether the paper provides sufficient details about the models/algorithms, datasets, and evaluation. Please take the authors' answers to the reproducibility checklist into account."

As can be seen, the authors are asked to fill the reproducibility checklist and to provide it together with the submission[20]. The checklist is available online[21]. It is divided into four main categories, each category comprising several items that can be checked or not depending on what is reported in the paper:

- **Models and algorithms**
  Items mainly concern : i) the description of the model/algorithms and their underlying assumption; ii) which software framework and version have been used when implementing.
- **Datasets**
  The items concern the description of the datasets and possibly corresponding citations and link to download the data.
- **Code**
  Items concern the code (both training and evaluation), its documentation (including dependencies) and the trained models.

[20] It seems that for MICCAI 2024, the authors did not have to fill the reproducibility checklist. We do not know what was the rationale for this decision.
[21] miccai2021.org/files/downloads/MICCAI2021-Reproducibility-Checklist.pdf

- **Experimental results**
  This section comprises all aspects regarding experimental results including: hyperparameters, data splits, evaluation metrics, assessment of variation (error bars), statistical testing, runtime, memory footprint, analysis of failures and discussion of clinical significance.

4.3 Let's try it!

The *Reproducibility Tutorial*[22] at MICCAI 2023 includes, among other things, a set of "fake papers" with varying degrees of reproducibility. The goal is to analyze the reproducibility of the papers based on a shortened version of the checklist and the information provided in the paper and in the repository. For most of the items there was no ground truth correction, and the participants had to choose a value between 1 (Not reproducible) and 10 (Reproducible). Then, they had to explain their choice, especially the ones at the extremities of the spectrum. We believe it is a good way to realize that assessing reproducibility is not a trivial task and that reproducibility comes as shades of gray. All the materials can be found under "Exercise 1 - Critical review tutorial" of the website[23].

In addition, the tutorial provides a second exercise where the aim is to perform a replication study ("Exercise 2 - Reproduce It Yourself!")[24]. The goal of this exercise is to give a concrete example of how to achieve exact reproducibility.

4.4 Analysis of reproducibility reviews at MICCAI 2023

We performed an analysis of the reproducibility section of the reviews of papers published at MICCAI 2023. The data and code necessary to reproduce this analysis is available as open source[25].

4.4.1. Methods

Reviews were automatically extracted from the MICCAI 2023 website[26] into a csv file. For each paper, we kept only the first three reviews, in order to have a constant number of reviews per paper and thus an homogeneous analysis across papers. We do not believe this leads to a loss of generality. The section of each review concerning reproducibility (referred to as *reproducibility review*

[22] miccai2023-reproducibility-tutorial.github.io
[23] miccai2023-reproducibility-tutorial.github.io/#exercise1
[24] miccai2023-reproducibility-tutorial.github.io/#exercise2
[25] github.com/reproducibility-reviews/reproducibility-reviews
[26] conferences.miccai.org/2023/papers

in the following, or simply as *review* when the context is unambiguous) was isolated and was analysed by human raters.

Two human raters (OC and ETS) each with more than 5 years of experience in MIC analysed the reproducibility reviews. To that purpose, they elaborated rating guidelines as follows. They first defined an initial version of the guidelines (v1). They then independently rated the first 40 papers (i.e. 120 reviews)[27]. They analysed discrepancies between ratings and revised the guidelines to make them more precise (v2). They independently rated a second time the first 40 papers. Again, they analysed the discrepancies and revised the guidelines (v3). This resulted in the final rating guidelines which are available in the repository[28]. In brief, the raters had to rate, for each reproducibility review, the following items:

- **Categories of the checklist**. Raters were requested to score "1" if the reviewer had made a comment that relates to at least one of the items falling into a category of the checklist. As a reminder, the categories are as follows:
  - **Models and algorithms**
  - **Datasets**
  - **Code**
  - **Experimental results**
- **Error bars and/or statistical significance**. Raters were requested to score "1" if the reviewer had commented on at least one of these two items.
- **Statement**. The raters had to indicate whether the reviewer had made an explicitly statement regarding the reproducibility of the paper in general and whether the statement was positive (+), negative (−) or unusable. Here are some examples of statements: "The work is reproducible" (+), "The reproducibility can be rated as sufficient" (+), "The work would be hardly reproducible" (−). Raters adopted a liberal view of statements. For instance, "seems ok" was considered a positive statement. Of course, this does not mean that the raters consider this to be a good quality review.
- **Comments**. The raters had to indicate whether the reviewer had made comments and whether these comments were all positive (+), all negative (−), a mixture of positive and negative comments (−/+) or unusable. Unlike statements, comments are not a general assessment but rather concern some specific aspects of reproducibility. Here are some examples of comments: "Methods were clearly described and code will be released"

[27] Papers were considered by alphabetical order of the title. We do not believe it introduces any bias in their selection.

[28] github.com/reproducibility-reviews/reproducibility-reviews/blob/main/human_rating/rating_guidelines.pdf

(+), "No code provided" (−) , "The hyper parameters are specified but there is no mention of a public repository anywhere" (−/+), "Through Implementation Details" (Unusable)[29].

- **Meta-category**. Using the statement and comments, we computed a meta-category which indicates whether the review is overall positive (+), negative (−) or unusable. To that purpose, we considered that statements took precedence over comments. Thus, if a statement was present, the value of the meta-category was equal to that of the statement. If there was no (or an unusable) statement, the meta-category was considered positive (+) if comments were positive, negative (−) if comments were negative or mixed and unusable if the comments were unusable.
- **Code is or will be available**. Raters had to put "1" if the reviewer had indicated that the code was provided at the time of reviewing or that the authors said they would release the code upon acceptance.
- **Code repository provided and not empty**. Raters had to put "1" if there was a code repository associated to the paper and the repository was not empty. Note that this is not based on the review but on the link associated to the paper on the MICCAI website. Also, note that the raters did not check the content of the code and thus cannot tell whether the code can be run or if it does what the paper describes.

The raters then independently rated the next 90 papers (i.e. 270 reviews), from 41 to 130. These ratings were used to assess inter-rater reliability as well as to produce the results of the analysis. We randomly chose to present the ratings of the first rater in this paper. However, we checked that results are very similar when using ratings of the second rater and the corresponding data is available in the repository[30].

For each item, inter-rater reliability was assessed using Cohen's κ [48]. For the analysis of reviews, the reporting frequency for each binary item was computed. For each paper, agreement between the three reviewers in terms of statement and meta-category was computed using Fleiss' κ [49], which generalizes Cohen's to more than two observers. All confidence intervals were computed using bootstrap with 1000 resamplings [50].

4.4.2. Results: inter-rater reliability

Inter-rater reliability ranged from substantial ($\kappa = 0.74$) to perfect ($\kappa = 1$) agreement. Results are presented in Table 3.

[29] It is hard to believe but yes, the review only contains these three words.
[30] github.com/reproducibility-reviews/reproducibility-reviews

Table 3: **Inter-rater reliability**. For each item, the table displays Cohen's κ and the associated 95% confidence interval (CI) computed using bootstrap.

	κ	95% CI
Models and algorithms	0.75	[0.66, 0.84]
Datasets	0.91	[0.85, 0.96]
Code	0.91	[0.86, 0.96]
Experimental results	0.86	[0.79, 0.93]
Error bars and/or statistical significance	1.00	[1.00, 1.00]
Statement	0.74	[0.67, 0.80]
Comments	0.82	[0.75, 0.87]
Meta-category	0.80	[0.73, 0.86]
Code is or will be available	0.87	[0.80, 0.93]
Code repository provided and not empty	1.00	[1.00, 1.00]

4.4.3. Results: reproducibility reviews

The main findings are summarized in Boxes 2 and 3.

The reproducibility section of the reviews was in general extremely short (median length: 16 words; inter-quartile range [IQR]: 18 words). For readers to get a more concrete idea of the amount of information it represents, Table 4 displays some randomly picked reviews which length falls between the 40^{th} and 60^{th} centile. Note that 30% of reviews contain less than 10 words (see some randomly chosen examples in Table 5). The full descriptive statistics of the different sections of the reviews is provided in Table S1 and the corresponding histograms are available here[31].

The checklist category that was most often commented upon was "Code" (47% of reviews (126/270), 95%CI [41%, 53%]). The lowest frequency was found for the "Experimental results" category (26% of reviews (69/270), 95%CI [20%, 31%]). Around 2% of reviews commented upon "Error bars and/or statistical significance". Full results are presented in Table 6.

61% of reviews provided a statement (164/270, 95%CI [55%, 66%]) and 72% came with comments (195/270, 95%CI [67%, 78%])[32]. Of note, 39% of reviewers which had made a positive statement regarding the reproducibility (indicating that they found that the reproducibility of the paper was overall satisfactory) provided no comments to substantiate their statement (52/132, 95%CI [32%, 48%]). Full results are presented in Tables S2, S3, S4, and S5.

[31] github.com/reproducibility-reviews/reproducibility-reviews/blob/main/supplementary_figures.pdf
[32] "Unusable" statements and comments are not taken into account.

Box 2: Reproducibility reviews at MICCAI 2023

- **No agreement between reviewers**
  - Fleiss'κ close to zero (from -0.05 to 0.05).
- **Very little information**
  - Extremely short reviews: the median number of words is 16. 30% have less than 10 words.
- **It should not be all about code**
  - Checklist category that was most reported about: "Code" (46%)
  - Item checklist that was least reported about: "Experimental results" (26%)
- **Statistical reproducibility: is there anybody out there?**
  - Less than 2% of reviewers reported about "Error bars and/or statistical significance"
- **Statements are often unsubstantiated**
  - 39% of reviewers who had made a positive statement provided no comments and thereby did not substantiate their statement in any way.

Box 3: Where is this code that you promised?

- **Code promised at submission but...**
  - For 53% of papers for which code was promised at submission (according to review comments), the code was actually missing for the published version (no link, broken link or empty repository).
- **Don't be sloppy**
  - For 30% of papers which provided a link for the code, the link was broken or the repository was empty.

Agreement between reviewers in terms of statement, comments and meta-category is reported in Table 7. Importantly, there was no agreement between reviewers with respective Fleiss'κ values of -0.05 (95%CI $[-0.14\%, 0.03\%]$) for statement and 0.02 (95%CI $[-0.09\%, 0.12\%]$) for meta-category.

For 87% of papers, at least one of the reviewers said that the code was or will be available (78/90, 95%CI $[79\%, 93\%]$). However, for 53% of these, the code was actually missing in the published version (no link, broken link or empty repository) (41/78, 95%CI $[41\%, 64\%]$). Finally, 68% of published papers provided an associated repository for the code (61/90, 95%CI $[58\%, 77\%]$). However, for 30% of these, the link was broken or the repository was empty (18/61, 95%CI $[18\%, 43\%]$). Details are provided in Table 8.

4.4.4. Limitations

The above analysis has several limitations. First, even though the rating of the reviews has good reproducibility, there remains a part of subjectivity related to the guidelines upon which the two raters agreed. The choices that were made for the guidelines are obviously debatable. Another limitation is that the estimation of "unkept promises" for the code uses the reviewers assessments which, in general, were found to be unreliable. It is thus true that the prevalence of "unkept promises" may be difficult to estimate. However, the finding "for 30% of the papers which provided a link for the code, the repository was empty or the link broken" does not depend on the reviewers' assessment and is reliable. We thus believe that it is reasonable to conclude that a substantial proportion of authors had a sloppy behavior regarding code release.

5. Where do we stand?

Some interesting measures and propositions... The MICCAI conference has put in place some interesting measures to assess and thus potentially promote reproducibility, based in particular on ideas from the 2020 Hackathon [16]. To the best of our knowledge, these measures are unique among MIC conferences and journals and this should be commended. More generally, awareness has risen in the MIC community as demonstrated through several publications (e.g. [1, 15, 51, 52]) and software efforts (e.g. [19, 53, 54]).

...but their full potential is not used However, much remains to be accomplished as we detail in the paragraphs below.

Reproducibility reviews are unreliable. Our analysis of MICCAI reviews has demonstrated that there is no agreement between reviewers. While reviewing

Table 4: 10 random reviews between 40 (13 words) and 60 (20 words) centiles

Word count	Review
15	"A public database, the OsteoArthritis Initiative is used for the validation of the proposed approach."
18	"The model parameters and experimental settings are provided. Code will be released. Dataset will be available on request."
15	"The results are on publicly available datasets and the authors promised to share the code."
19	"The clinical dataset is not open publicly available. It prevents the reproducibility assessment of the impact in clinical setting"
14	"I consider the paper is reproducible as the code is released at anonymous Github."
18	"The reproducibility of the method seems to be easily done, the authors provide enough details to do so."
15	"The link of the source code is given in the the Implementation details (Section 3)."
19	"Due to the complexity of the architecture, the information described may not be sufficient to easily reproduce the results."
15	"The model and datasets are well described, but they provide no code nor raw data."
18	"The authors provide detailed information on their methodology, train/test procedure, and used datasets, resulting in a reproducible work."

Table 5: 10 random reviews between 0 (1 word) and 30 (11 words) centiles

Word count	Review
2	"Good reproducibility"
4	"The method is reproducible."
9	"The paper provided the source code of proposed model."
4	"Appears to be reproducible."
9	"The authors claim to open-source the code and data."
7	"This method might be easy to reproduce."
7	"This paper is based on public dataset."
3	"Reproducibility is guaranteed."
4	"Not enough for reproduction."
8	"The reproducibility of this paper is not clear."

is always partly subjective, such an extreme situation is abnormal. One would expect to have at least some moderate agreement between reviewers on a matter for which some objective criteria can be agreed upon. Some of the reviews are even in complete contradiction. Please refer to Table 9 for some examples which, although cherry-picked, are in line with our systematic assessment with a κ around zero for agreement between reviewers.

Reproducibility reviews provide little information. Reviews are extremely short (median is 16 words). Reviewers' statements are often unsubstantiated leaving the area chair to wonder whether the reviewer has really checked the reproducibility of the paper. The most common type of comment was regarding the code. While it is great to encourage open source, it is a sign that the whole concept of reproducibility is not correctly understood by many reviewers since it goes way beyond code. In particular, there were few comments on experimental results and almost none (2% of reviews) on statistical aspects (error bars and/or statistical significance). As a result, reviews do not provide useful information for area chairs and thus it is difficult to see how reproducibility can be taken into account in the decision process. Also, only few reviews provide useful feedback to the authors. This is a missed opportunity to contribute to increase reproducibility in the community.

Is reproducibility taken seriously? There are several signs that makes one wonder whether reproducibility is taken seriously. One is the lack of reliability and informativeness of reviews. Another sign, probably more worrying, is that many authors do not keep their promises regarding code availability (e.g. 30%

Table 6: **Frequency of comments on checklist categories.** For each of the four categories of the checklist (Models and algorithms, Datasets, Code, Experimental results), we report the proportion of reviews which mention at least one item of the category. For "Error bars and/or statistical significance", we report the proportion of reviews which mention at least one of these two items. 95% confidence intervals (CI) are computed using bootstrap.

Category/item	%	95% CI	N
Models and algorithms	29%	[24%, 35%]	78/270
Datasets	33%	[28%, 39%]	90/270
Code	47%	[41%, 53%]	126/270
Experimental results	26%	[20%, 31%]	69/270
Error bars and/or statistical significance	2%	[0%, 4%]	5/270

Table 7: **Agreement between reviewers.** We report Fleiss'κ to measure the agreement between reviewers for "Statement" and "Meta-categories". As one can see, there is no agreement between reviewers, with κ values close to zero. 95% confidence intervals (CI) are computed using bootstrap.

	κ	95% CI
Statement	-0.05	[-0.14, 0.03]
Meta-categories	0.02	[-0.09, 0.12]

Table 8: **Was the code promised delivered?** Columns indicate whether at least one of the three reviewers has indicated that the code was available or will be upon acceptance ("Yes") or not ("No info"). Rows indicate whether the published paper came was a valid repository, an invalid repository (broken link or empty repository) or no repository.

Code promised (according to reviewers) → Code available (after acceptance) ↓	No info	Yes	Total
No repo	6 (7%)	37 (41%)	43 (48%)
Invalid repo	2 (2%)	16 (18%)	18 (20%)
Valid repo	4 (4%)	25 (28%)	29 (32%)
Total	12 (13%)	78 (87%)	90 (100%)

Table 9: Some cherry-picked papers for which reviews are in complete disagreement. Note that, even though the papers below were cherry-picked, the fact that reviewers disagree is backed-up by scientific evidence (see Table 7).

Paper 1

- **Reviewer 1**
  - ➤ "The present application-oriented study has poor reproducibility based since:
    - data is not made available code of the whole pipeline is not made to, only certain parts like the standard nnUnet
    - trained models are not made available
    - computational requirements were not described
    - due to small number of test cases, reproducing the results on different dataset could be challenging (especially since the test data structure is not fully reported, i.e. sex, gestational age, diagnosis, ethnicity, etc.)"
- **Reviewer 2**
  - ➤ "The paper has provided enough details for ensuring general reproducibility."
- **Reviewer 3**
  - ➤ "Information to reproduce the experiments is provided."

Paper 2

- **Reviewer 1**
  - ➤ "This is a 100% reproducible."
- **Reviewer 2**
  - ➤ "Reproducible. Method details described well."
- **Reviewer 3**
  - ➤ "The paper is unlikely to be reproduccible with the current information provided."

Paper 3

- **Reviewer 1**
  - ➤ "The authors provide links to the two datasets used in their study and mention several hyperparameters used for their model and baselines. However, some important hyperparameters are not explicitly stated, such as the learning rate, batch size, and optimizer. The authors do note that their code will be made publicly available, which will allow interested readers to access this information."
- **Reviewer 2**
  - ➤ "The reproducibility checklist agrees to what can be seen in the paper."
- **Reviewer 3**
  - ➤ "It's easy to reproduce."

of empty/broken link repositories): faking open science is much worse than doing closed science. Also, even though we could not assess this systematically, in our experience as reviewers, it is not uncommon that there are major discrepancies between the checklist and the paper (some authors even tick all boxes regardless of the paper content).

Potential underlying causes. It seems that there is a lack of understanding of reproducibility by the community in general. This is particularly obvious based on the analysis of the reviews. Review guidelines could certainly be more detailed but this is unlikely to be the only explanation. More likely, the reviews only reflect a lack of training of the community as a whole.

6. Where could we go?

This chapter concludes with outlining some possible avenues for progress for the MIC community. Of course, these are only suggestions to stimulate the discussion.

Best practices guidelines. There is a clear need for *consensus* guidelines as already suggested in [16]. Such consensus guidelines have recently been published for validation metrics [38]. We need similar guidelines for evaluation procedures and statistical analyses. In other words, we need not only to know what metrics we should compute but also how we should estimate them. We believe that we also need another set of consensus guidelines for reproducibility as a whole, that would in particular cover the different building blocks described in Section 3. Very valuable sources of information already exist but guidelines which are based on a *consensus* and which are *adapted* to MIC are needed. These guidelines would be key components of a more general effort aiming at training the whole community. Indeed, part of our paper was focused on the review process but it should be obvious that scientists should be trained to reproducibility not only as reviewers but, even more importantly, as authors.

Improving the review process. This is also critical as it is currently not informative. A first step could be to provide more precise guidelines to the reviewers. This could explain what components are expected and what is typically considered a "good review". Box 4 suggests some ideas that could serve to that purpose, while Box 5 provides some tentative examples of good reproducibility reviews. Even though this is more minor, the checklist should probably be revised and possibly shortened (with the hope that it will more often be properly filled). Please note that it seems that for MICCAI 2024, the

Box 4: Tentative guidelines for reproducibility reviews

Here, we provide tentative guidelines for high-quality reproducibility reviews. Of course, these would need to be discussed within the community to produce *consensus guidelines*. Please refer to Box 5 for examples of what a good review could look like.

- **A clear statement adapted to the type of paper and its claims**
  - ➤ Start your review with one (or a few) sentences stating your general opinion about the reproducibility of the paper. Adapt your statement to the content of the paper and to its claims. For example: i) if a paper uses a subset of a public dataset, it should clearly say which criteria were used to select the samples and provide their list; ii) it is okay that a paper does not release the data if there are some legal constraints that prevent it; iii) a paper that claims that the proposed method outperformed the baseline needs to provide statistical evidence.
- **Back-up your statement with comments**
  - ➤ Always provide at least one comment for each of the four categories of the checklist (models and algorithms, datasets, code, experimental results), even if you believe the reproducibility is perfect: this will tell the area chair that you have carefully taken all aspects into account.
  - ➤ Use the comments to list all the strengths and weaknesses that relate to reproducibility. This provides useful information to the area chair and valuable feedback to the authors.
  - ➤ Do not overlook experimental aspects (including data splits, validation metrics, descriptive and inferential statistics). Always have in mind that MIC is an experimental science. Reproducibility is not only about open code and open data.
- **Regarding the checklist**
  - ➤ Provide one (or a few) sentence saying whether the checklist filled by the authors reflects the content of the paper[a].

[a]Ideally, this could be replaced by the reviewer marking whether each item of the checklist was correctly filled, provided that this can be conveniently implemented in the review website.

Box 5: Tentative examples of good reproducibility reviews

Here, we provide tentative examples of high-quality reproducibility reviews that follow the guidelines proposed in Box 4. Of course, the reviews look generic because they are not related to a specific paper. Real reviews would contain more specific elements.

- **Example 1: you think everything is perfect.**

  ➤ "The reproducibility of this work appears excellent.
  The authors have clearly described the methods, the dataset is public and the data selection is clear, the code is available and documented, the experimental part is thorough including a clear description of the data splits and descriptive and inferential statistics.
  The checklist filled by the authors matches what is reported in the paper."

- **Example 2: you think reproducibility is adequate given the type of paper, even though exact reproducibility would not be possible.**

  ➤ "The reproducibility of this work appears adequate with respect to the constraints. Note that the data were not shared because the authors do not have the authorization to do so: this will prevent exact reproducibility but is perfectly acceptable.
  The authors have clearly described the methods. As said above, data cannot be shared but it is clearly described with all necessary information. The code is available and documented. The experimental part contains all the necessary information.
  The checklist filled by the authors matches what is reported in the paper."

- **Example 3: mixed review, there are some positive aspects but also some important weaknesses.**

  ➤ "This work has a moderate level of reproducibility with some important weaknesses regarding the experimental part.
  The authors have clearly described the core of the model and the corresponding implementation is available. On the contrary, preprocessing is not clearly described and the corresponding code not provided. The data are public and the data selection is clear. The experimental part only reports point estimates without descriptive (e.g. standard deviation) nor inferential statistics (e.g. confidence intervals).
  The checklist filled by the authors matches what is reported in the paper."

- **Example 4: the authors ticked all items of the checklist but the reproducibility is actually very poor.**

  ➤ "This work has a low level of reproducibility unlike what is stated in the checklist. The description of the method is decent although it would not be possible to reproduce it based only on the paper. The data are not shared. The clinical characterization of the data is insufficient. The experiments are loosely described: apparently there is no independent test set, it is unclear how the hyperparameters were set, the authors report only mean values without standard deviation nor confidence intervals.
  The authors have ticked all the boxes in the checklist which is inconsistent with what is reported in the paper."

authors did not need to fill the checklist. We do not know the rationale for this decision but, at first glance, it looks like it goes in the wrong direction. The checklist was not perfect and could be improved but we believe that removing it is not a good signal to encourage reproducibility. We also suggest to add a grade on reproducibility in addition to comments. More generally, the idea of a reproducibility chair, as proposed by the 2020 Hackathon [16], appears very useful to coordinate all this.

Tracking progress. It would be nice to put in place tools that allow tracking progress throughout years. For instance, if this can be implemented in the review system, it would be useful to ask reviewers to check the checklist: putting checkmarks along those of the authors to indicate whether the authors' checklist adequately reflects the paper. This would allow not only quantitatively assessing whether authors correctly fill the checklist but also identifying what are the most common practices in terms of reproducibility. In a similar spirit, it would be useful to have checkboxes for reviewers to indicate whether authors have promised to release code and/or data upon acceptance. This would allow checking which proportion of authors keep their promises. Finally, one could consider publishing the checklists along with the papers and reviews after acceptance.

Incentives for being reproducible. As suggested by Balsinger et al [16], there is a need to promote reproducibility efforts made by authors. At the moment, it is unclear how reproducibility is taken into account for acceptance. We believe it should be but that the requirements should be adapted to the types and the claims of papers: this requires training of the reviewers to the concepts of reproducibility. The suggestion of making checklists public that we made above can also contribute to promote efforts. An official list of papers which release code and/or data would also be useful as currently only unofficial lists have been created [33,34]. Moreover, it would be nice to have more papers unveiling reproducibility issues (in the broad sense including not only replication studies but also identification of potential biases for instance). This may require to have a dedicated topic and dedicated area chairs. A reproducibility award could be considered even though its scope and criteria remain to be determined.

Expand efforts beyond MICCAI. Many of our comments so far have concerned MICCAI: the only reason is that it is, to our knowledge, the only publishing venue of our community which has taken actions to promote reproducibility. But of course, we are not pointing fingers at MICCAI. On the contrary, MICCAI should be commended for implementing those measures.

[33] github.com/JunMa11/MICCAI-OpenSourcePapers
[34] github.com/yiqings/MICCAI2022_paper_with_code

More generally, efforts should not only concern MICCAI but also other conferences and journals of the community. We are well aware that this is not an easy task: it may even be that some web-based review systems make it impossible at the moment. For instance, one of the authors (OC) is co-chair of the Image Processing conference at the SPIE Medical Imaging symposium and would not be able to make such changes to the review process which is common to all conferences of the symposium and which rely on an electronic submission/review system over which conference chairs do not have control. Finally, we think that it is even more important to implement measures for reproducibility in journals than at conferences. Journal papers are what should contain fully matured and consolidated research. Moreover, it is understandable that a conference paper is rushed while a journal paper has less excuse for escaping reproducibility.

Let's progress together. "Criticism is easy, and art is difficult" as French dramatist Destouches wrote. We would thus like to make clear that we include ourselves in this criticism and the reader will easily find papers by us which miss several reproducibility building blocks (e.g. [55, 56, 57]). More specifically, our paper on convolutional neural networks for classification of Alzheimer's disease [52] has adequate exact reproducibility but falls short on statistical aspects. If we had to write it today, the statistical part would be different. Therefore, we need to progress together as a community. As we will argue below, we should not expect such progress to be uniform nor have the same reproducibility objectives for all studies.

One size does not fit all. In the above, we have suggested the need for guidelines and review processes regarding reproducibility. We would like to make clear that we do not advocate for a dogmatic view on that matter, quite the opposite. As explained in this paper, there are different types of reproducibility. There will never be a "one-size-fits-all" format that is adapted to all MIC papers. For instance, asking for exact reproducibility for all papers would be complete non-sense. This would lead researchers to focus on public datasets while we also need work on clinical routine data which are much more difficult to share. This in turn would lead to lower the conceptual reproducibility and ultimately the clinical translation. On the contrary, it is crucial to realize that reproducibility requirements need to be adapted to the type and content of the paper.

Acknowledgments

The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part

of the "Investissements d'avenir" program, ref. ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and ANR-10-IAIHU-0006 (Agence Nationale de la Recherche - 10-IA Institut Hospitalo-Universitaire-6). The authors are grateful to Nicolas Gensollen for the code review.

References

[1] Varoquaux G, Cheplygina V (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digital Medicine 5(1):1–8

[2] Wagner SJ, Matek C, Shetab Boushehri S, Boxberg M, Lamm L, Sadafi A, Waibel DJ, Marr C, Peng T (2022) Make deep learning algorithms in computational pathology more reproducible and reusable. Nature Medicine 28(9):1744–1746

[3] Baker M (2016) 1,500 scientists lift the lid on reproducibility. Nature 533(7604)

[4] Gundersen OE (2020) The reproducibility crisis is real. AI Magazine 41(3):103–106

[5] Begley CG, Ioannidis JP (2015) Reproducibility in science: improving the standard for basic and preclinical research. Circulation research 116(1):116–126

[6] Collaboration OS (2015) Estimating the reproducibility of psychological science. Science 349(6251):aac4716

[7] Begley CG (2013) An unappreciated challenge to oncology drug discovery: pitfalls in preclinical research. American Society of Clinical Oncology Educational Book 33(1):466–468

[8] Stupple A, Singerman D, Celi LA (2019) The reproducibility crisis in the age of digital medicine. NPJ Digital Medicine 2(1):1–3

[9] Gundersen OE, Kjensmo S (2018) State of the art: Reproducibility in artificial intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32

[10] Hutson M (2018) Artificial intelligence faces reproducibility crisis. Science

[11] 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

[12] 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. Journal of Machine Learning Research 22

[13] McDermott M, Wang S, Marinsek N, Ranganath R, Ghassemi M, Foschini L (2019) Reproducibility in machine learning for health. arXiv preprint arXiv:190701463

[14] Beam AL, Manrai AK, Ghassemi M (2020) Challenges to the reproducibility of machine learning models in health care. JAMA 323(4):305–306

[15] Langer SG, Shih G, Nagy P, Landman BA (2018) Collaborative and reproducible research: goals, challenges, and strategies. Journal of Digital Imaging 31(3):275–282

[16] 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

[17] Colliot O, Thibeau-Sutre E, Burgos N (2023) Reproducibility in machine learning for medical imaging. In: Machine Learning for Brain Disorders, Springer, chap 21, pp 631–653

[18] Varoquaux G, Colliot O (2023) Evaluating machine learning models and their diagnostic value. Machine Learning for Brain Disorders pp 601–630

[19] 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. Computer Methods and Programs in Biomedicine 220:106818

[20] Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Sciences 22:1359–1366

[21] 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. Science Translational Medicine 13(586):eabb1655

[22] 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. Nature Neuroscience 20(3):299–303

[23] Goodman SN, Fanelli D, Ioannidis JP (2016) What does research reproducibility mean? Science Translational Medicine 8(341):341ps12–341ps12

[24] Plesser HE (2018) Reproducibility vs. replicability: a brief history of a confused terminology. Frontiers in Neuroinformatics 11:76

[25] Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC (2021) Reproducibility standards for machine learning in the life sciences. Nature Methods 18(10):1132–1135

[26] Gorgolewski KJ, Poldrack RA (2016) A practical guide for improving transparency and reproducibility in neuroimaging research. PLoS Biology 14(7):e1002506

[27] 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

[28] Lemieux L, Hagemann G, Krakow K, Woermann FG (1999) Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 42(1):127–135

[29] Laurienti PJ, Field AS, Burdette JH, Maldjian JA, Yen YF, Moody DM (2002) Dietary caffeine consumption modulates fMRI measures. NeuroImage 17(2):751–757

[30] Borga M, Ahlgren A, Romu T, Widholm P, Dahlqvist Leinhard O, West J (2020) Reproducibility and repeatability of MRI-based body composition analysis. Magnetic Resonance in Medicine 84(6):3146–3156

[31] 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. European Radiology 30(1):195–205

[32] 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 transactions on medical imaging 17(3):463–468

[33] 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

[34] Loizillon S, Bottani S, Maire A, Ströer S, Dormont D, Colliot O, Burgos N (2023) Automatic motion artefact detection in brain t1-weighted magnetic resonance images from a clinical data warehouse using synthetic data. Medical Image Analysis p 103073

[35] 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. Nature Reviews Neuroscience 18(2):115–126

[36] 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 p 119623

[37] Turkyilmaz-van der Velden Y, Dintzner N, Teperek M (2020) Reproducibility starts from you today. Patterns 1(6):100099

[38] Maier-Hein L, Reinke A, Christodoulou E, Glocker B, Godau P, Isensee F, Kleesiek J, Kozubek M, Reyes M, Riegler MA, et al (2022) Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv preprint arXiv:220601653

[39] 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

[40] 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. Scientific Data 3(1):1–9

[41] 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

[42] Bengio Y, Grandvalet Y (2003) No unbiased estimator of the variance of k-fold cross-validation. Advances in Neural Information Processing Systems 16

[43] Nadeau C, Bengio Y (1999) Inference for the generalization error. Advances in neural information processing systems 12

[44] Hothorn T, Leisch F, Zeileis A, Hornik K (2005) The design and analysis of benchmark experiments. Journal of Computational and Graphical Statistics 14(3):675–699

[45] Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of p-hacking in science. PLoS biology 13(3):e1002106

[46] Bottani S, Burgos N, Maire A, Saracino D, Ströer S, Dormont D, Colliot O (2023) Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse. Medical Image Analysis 89:102903

[47] 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. Investigative Radiology 53(11):647

[48] Cohen J (1960) A coefficient of agreement for nominal scales. Educational and psychological measurement 20(1):37–46

[49] Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological bulletin 76(5):378

[50] Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. CRC press

[51] 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

[52] 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. Medical Image Analysis 63:101694

[53] Glatard T, Lartizien C, Gibaud B, Da Silva RF, Forestier G, Cervenansky F, Alessandrini M, Benoit-Cattin H, Bernard O, Camarasu-Pop S, et al (2012) A virtual imaging platform for multi-modality medical image simulation. IEEE transactions on medical imaging 32(1):110–118

[54] Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, Murrey B, Myronenko A, Zhao C, Yang D, et al (2022) Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:221102701

[55] Colliot O, Mansi T, Bernasconi N, Naessens V, Klironomos D, Bernasconi A (2006) Segmentation of focal cortical dysplasia lesions on MRI using level set evolution. Neuroimage 32(4):1621–1630

[56] Colliot O, Chételat G, Chupin M, Desgranges B, Magnin B, Benali H, Dubois B, Garnero L, Eustache F, Lehéricy S (2008) Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. Radiology 248(1):194–201

[57] Burgos N, Cardoso MJ, Samper-González J, Habert MO, Durrleman S, Ourselin S, Colliot O (2021) Anomaly detection for the individual analysis of brain PET images. Journal of Medical Imaging 8(2):024003–024003

Supplementary Material

Table S1: **Number of words for the different sections of the reviews**.

Category	Mean	Min	Max	Median	σ	IQR
Contribution	60.90	0	167	56	30.56	42
Strengths	60.15	0	433	49	45.96	46
Weaknesses	97.22	0	494	72	83.84	97
Reproducibility	22.58	0	429	16	26.30	18
Detailed comments	109.97	0	1240	70	124.54	115
Justification	34.88	0	166	28	26.96	30
Total	385.71	0	2219	336	210.88	234

σ: standard-deviation, IQR: inter-quartile range.

Table S2: **Statements.** We report whether the reviewer has made a statement regarding the overall reproducibility of the paper and whether this statement was positive (+), negative (−) or unsusable. (none) indicates that no statement was made. As a reminder, there was no agreement between reviewers (Fleiss'$\kappa = -0.05$, 95%CI : $[-0.14, 0.04]$). Thus, one cannot draw conclusions about the actual reproducibility of the papers.

	%	95% CI	N
(+)	49%	[43%, 55%]	132/270
(−)	12%	[8%, 16%]	32/270
(none)	37%	[31%, 43%]	100/270
(unusable)	2%	[1%, 4%]	6/270

Table S3: **Comments.** We report whether the reviewer has made comments and whether these comments were positive (+), negative (−), mixed (−/+) or unsusable. (none) indicates that no comment was made.

	%	95% CI	N
(+)	31%	[25%, 36%]	83/270
(−)	19%	[14%, 23%]	50/270
(−/+)	23%	[18%, 28%]	62/270
(none)	22%	[17%, 27%]	60/270
(unusable)	6%	[3%, 9%]	15/270

Table S4: **Meta-category.** We report whether the meta-category, which describes the overall opinion of the reviewer by combining statements and comments, was positive (+), negative (−), or unsusable. As a reminder, there was no agreement between reviewers (Fleiss'$\kappa = 0.02$, 95%CI : $[-0.09, 0.13]$). Thus, one cannot draw conclusions about the actual reproducibility of the papers.

	%	95% CI	N
(+)	60%	[54%, 66%]	163/270
(−)	32%	[27%, 38%]	87/270
(unusable)	7%	[4%, 11%]	20/270

Table S5: **Statements vs comments.** We report the counts of couples (statement, comments) for each possible combination.

Comments → Statement ↓	(+)	(−)	(−/+)	(none)	(unusable)	Total
(+)	52	7	21	52	0	132
(−)	0	21	8	3	0	32
(none)	31	22	33	5	9	100
(unusable)	0	0	0	0	6	6
Total	83	50	62	60	15	270