

Article

Not peer-reviewed version

CALIFRAME: A Proposed Method of Calibrating Reporting Guidelines with FAIR Principles to Foster Reproducibility of AI Research in Medicine

[Kirubel Biruk Shiferaw](#)^{*}, Irina Balaur, Danielle Welter, [Dagmar Waltemath](#), Atinkut Alamirrew Zeleke

Posted Date: 23 July 2024

doi: 10.20944/preprints202407.1830.v1

Keywords: FAIR principles; Reporting guidelines; Medicine; Machine learning; Artificial intelligence; Calibration



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

CALIFRAME: A Proposed Method of Calibrating Reporting Guidelines with FAIR Principles to Foster Reproducibility of AI Research in Medicine

Kirubel Biruk Shiferaw ^{1,*}, Irina Balaur ², Danielle Welter ³, Dagmar Waltemath ¹
and Atinkut Alamirrew Zeleke ¹

¹ Department of Medical Informatics, Institute for Community Medicine, University Medicine Greifswald, Germany

² Luxembourg Centre for Systems Biology, University of Luxembourg, 6 Avenue du Swing, L-4367 Belvaux, Luxembourg

³ Luxembourg National Data Service, 6, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

* Correspondence: Kirubel Biruk Shiferaw, Walther-Rathenau-Str. 48, Department of Medical Informatics, University Medicine Greifswald, D-17475 Germany; E-mail: s-kishif@uni-greifswald.de.

Abstract: Procedural and reporting guidelines frame the process of scientific practices and communications among researchers and the community at large. In the pursuit of fostering reproducibility, several methodological frameworks have been proposed by several initiatives. Nevertheless, recent studies indicate that data leakage and reproducibility are still prominent challenges. Recent studies have shown the transforming potential of incorporating the FAIR (Findable, Accessible, Interoperable and Reusable) principles in the work-flow of different context such as software and machine learning model development stages to cultivate open science. In this work, we introduce a framework to calibrate reporting guidelines against the FAIR principles in order to foster reproducibility and open science. We adapted the “Best fit” framework synthesis approach to develop the calibration framework. We propose a series of defined workflows to calibrate reporting guidelines with FAIR principles and a use case to demonstrate the process. By integrating FAIR principles with established reporting guidelines, the proposed framework bridges the gap in accommodating both FAIR metrics and reporting guidelines and benefits from advantages of these major integrated components.

Keywords: FAIR principles; reporting guidelines; medicine; machine learning; artificial intelligence; calibration

Introduction

The dynamic landscape of Artificial Intelligence (AI) requires transparency, trustworthiness, and reproducibility of research outcomes [1]. In the pursuit of fostering reproducibility, several methodological frameworks have already been designed [2–4]. Procedural and reporting guidelines frame the process of scientific practices and communications among researchers and the community at large. Nevertheless, recent studies indicate that leakage and reproducibility in AI-based studies resulted in overoptimistic conclusions and sometimes “super heroic” presentations of AI models [5–7]. Good scientific practice demands that key steps in the data pre-processing, model development and curation, validation and deployment strategies should be reported [8].

Reporting guidelines for AI-related studies in the healthcare context should be designed to mitigate the aforementioned reporting gap by facilitating transparency and reproducibility. While it is unrealistic to aim for a gold standard guideline that works for all contexts including specific domains and study designs [9], recent studies have shown the transforming potential of incorporating the FAIR (Findable, Accessible, Interoperable and Reusable) principles [10] in the work-flow of AI model development and reporting stages [11,12].

In this paper, we introduce a framework to calibrate reporting guidelines against the FAIR principles, thereby enhancing reproducibility. The term “Calibration” in this context refers to a harmonization of reporting guidelines with the FAIR principles without altering the nature and content of the guidelines. The proposed framework resourcefully integrates established guidelines with the FAIR principles, leading to the creation of a calibrated reporting guideline that considers

domain-specific FAIR indicators [13] for AI research. Ultimately, the proposed framework presents a holistic solution that transcends disciplinary boundaries. We believe that the calibrated guidelines contribute to an improved culture of shared knowledge.

Methods

The framework for calibration was developed by adapting the “Best fit” framework synthesis approach [14]. This approach identifies and selects existing relevant frameworks and adapts or merges them to form a new framework applicable to the intended research purpose. Thus, the approach enables researchers to take advantages of the strengths of existing frameworks while tailoring them to accommodate the most recent advances in the field.

To refine and optimize our approaches, we engaged in iterative process of improvement, incorporating feedback/insights from preliminary analyses and regular communications. This iterative approach enabled us to continuously enhance the robustness, precision, and reliability of our proposed method of FAIR calibration.

Results

The result is a series of defined workflow steps for calibration and a use case to demonstrate its applicability. Calibrating reporting guidelines is imperative to develop a tailored solution, to improve efficiency (using already available sources to address the gap of one guideline with the other instead of developing *ab initio*) and to bridge the disciplinary gap by combining components from different guidelines. The work-flow to develop the calibrated reporting guideline is represented in Figure 1.

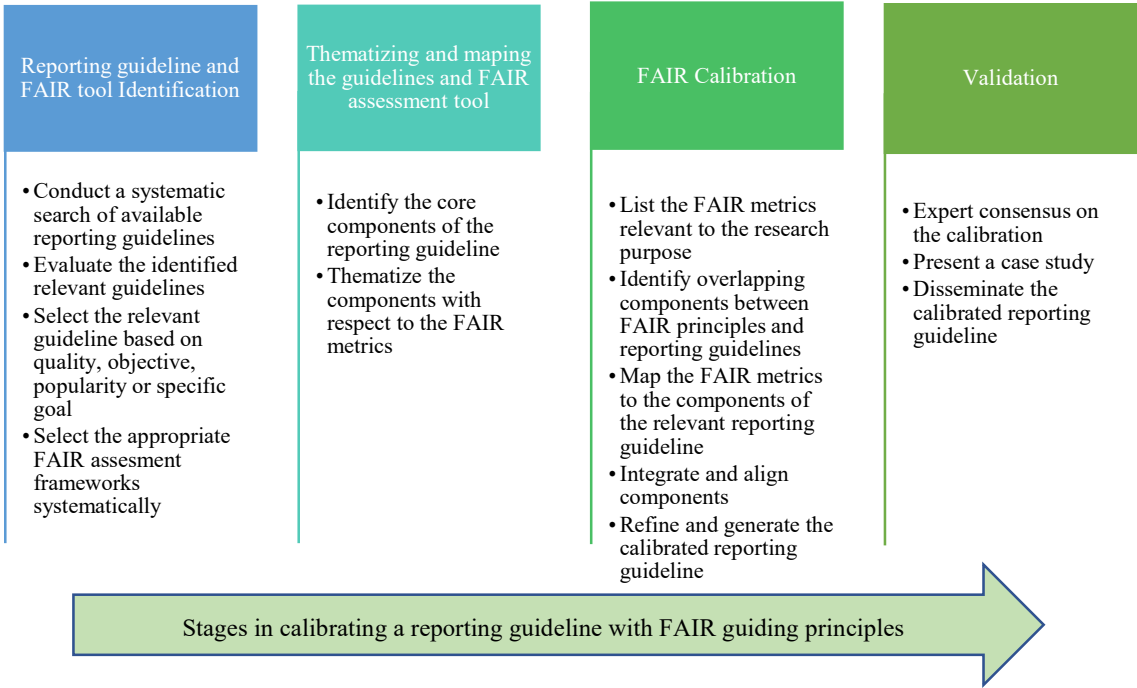


Figure 1. Diagrammatic representation of the stages for calibrating the existing reporting guidelines with the FAIR principles.

Stage 1: Identification of reporting guideline and FAIR assessment tool

The starting point in the calibration process is the reporting guideline, which is defined as a minimum checklist containing relevant items to validate studies’ readability, reproducibility and reliability [15]. If the reporting guideline is not already identified by the researcher, the identification process should involve a systematic search of the guidelines using appropriate keywords and databases/sources. After carefully selecting the available guidelines (based on the inclusion and exclusion criteria set by the researcher prior to the search), a comprehensive evaluation needs to be commenced to select the most appropriate guideline. The comparison of identified guidelines can be made in terms of quality, objective, popularity or specific goal. For example, the AGREE II (Appraisal

of Guidelines, Research and Evaluation) tool can be used to assess the quality of clinical practice guidelines [16,17]. It was developed by experts to evaluate methodological quality guidelines using six domains (scope and purpose, stakeholder involvement, rigour of development, clarity of presentation, applicability and editorial independence) of guideline quality assessment metrics [18]. The PRISMA flowchart is recommended to document the search results, reasons of inclusion and exclusion, number of studies and sources [19].

The FAIR assessment tool is another major component of this workflow. Several FAIR assessment metrics have been proposed by different FAIR initiatives striving to improve and assess the FAIRness of resources (<https://fairassist.org#!/>). Hence, multiple options for FAIR evaluation are available, and researchers should invest the required time and effort to select the most appropriate metrics for their research objectives.

Use case: We demonstrate the applicability of our framework with a concrete use case, generating a FAIR calibrated reporting guideline for clinical trials that involves interventions with an AI component. Initially, we conduct a systematic search and quality assessment to identify the most appropriate reporting guideline and FAIR assessment tool. For demonstration purposes, we selected the Consolidated Standards of Reporting Trials-Artificial Intelligence extension (CONSORT-AI) [20] guideline and the Research Data Alliance (RDA) FAIR Data Maturity Model [21]. CONSORT-AI is one of the widely used and high-quality reporting guidelines with 25 core items and 14 sub-specific items. The RDA FAIR Maturity Model evaluates compliance with each FAIR principle through one or more indicators. Each indicator is associated with an impact level (essential, important, or useful) and indicators target both project data and associated metadata [21]. The RDA FAIR Data Maturity Model describes 41 data and metadata indicators with detailed description in relation to the FAIR principles with details of how each indicator is assessed [21].

Stage 2: Thematizing and mapping the guideline

The chosen guideline should be separated into its key components such as title and abstract, introduction, methods, results, discussion, conclusion, funding, Supplementary Materials, appendices and references.

Similarly, the elements of the FAIR metric need to be broken down into the four core components: findability (F), accessibility (A), interoperability (I) and reusability (R). All the FAIR indicator/metrics should be listed together with a detailed description of (1) what is being assessed and (2) how it is measured.

Use case: At this stage, we identify the elements from both the reporting guideline corresponding to each principle. For CONSORT-AI, we list all the 51 items along with their descriptions and means of assessment. Similarly, we list all the metrics elements of RDA FAIR indicators along with their description and method of assessment.

Stage 3: FAIR Calibration

After clearly identifying the key components, the next step is the FAIR calibration, which refers to the systematic mapping of commonalities and complementarities between the FAIR principles and the identified reporting guideline. A core step here is to thoroughly evaluate the alignment of the selected reporting guidelines and the FAIR principles. The “Best Fit” framework synthesis method facilitates the evaluation of the alignment and the development of a new component to incorporate the non-aligning components [22]. To do so, a profound understanding of the FAIR metrics and the identified guideline is required.

Use case: After clearly identifying the elements of both RDA FAIR Indicators and CONSORT-AI items, we then identify commonalities and complementarities. For example, item 23 of CONSORT-AI smoothly align with Findability indicators in RDA FAIR indicators (F101M, F102M, F301M, F303M, F401M). Based on this evaluation, we suggest a solution to calibrate. It is also important to note that some items might not align or map to any FAIR indicators. In this case the item should be kept for the next stage of the calibration process.

Stage 4: Validation

To ensure the validity and effectiveness of the calibrated reporting guidelines, a rigorous validation process should be conducted. A panel of inter-disciplinary experts in AI research, ethics, and reproducibility should be convened. The guideline, along with its integrated FAIR principles, should be presented to the expert panel for review and amendments should be done in iterative manner. The experts should also evaluate the alignment of guidelines with FAIR principles, identify

potential conflicts, and suggest refinements. The refined and validated reporting guideline should be disseminated for further refinement and implementation. Figure 2 shows the process of calibrating reporting guidelines with FAIR principles.

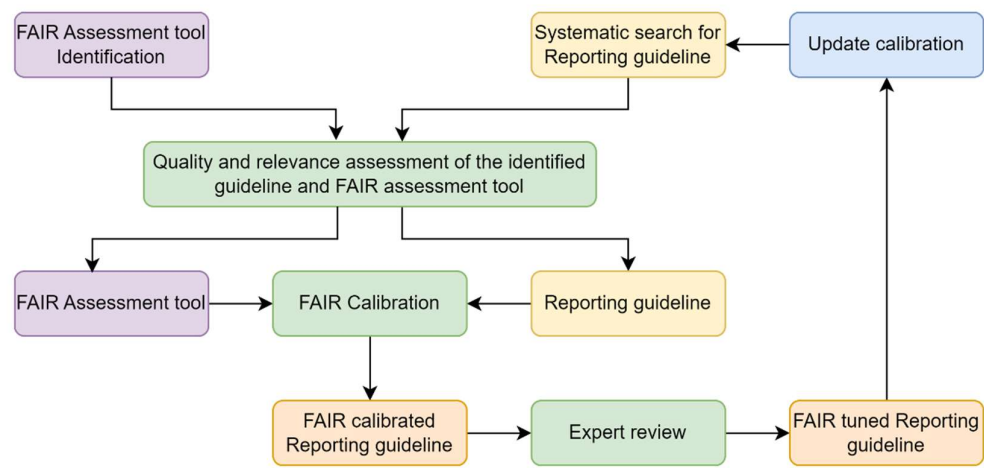


Figure 2. The FAIR calibration process of reporting guidelines. Identification of the FAIR assessment tool and reporting guideline can be performed in parallel. Similarly, the quality and relevance assessment of the identified guideline and FAIR assessment tool also can be done in parallel or one after the other. The colours differentiate FAIR assessment tool(purple), reporting guideline(yellow), process/activities(green) and the iterative update task (blue).

Use case: After having the FAIR calibrated reporting guideline, we then invite experts to comment on the proposed reporting guideline to make sure that elements of both the reporting guideline and FAIR principles are harmonized. Further, we plan disseminating the calibrated guideline through workshops, publication and other scientific communication to obtain additional suggestions and to achieve community consensus.

Discussion

The FAIR guiding principles presents a broad scheme that aims to make data and metadata findable, accessible, interoperable and reusable by both humans and machines [23]. It plays a substantial role in the path to effective data stewardship. In this age of information abundance, embracing the FAIR principles is not merely a choice, but a necessity, as they empower us to shape our data-driven aspirations into a vivid reality of innovation and progress.

Different approaches of integrating the FAIR principles in reporting AI interventions have been proposed by researchers in various domains [12]. Mobilizing FAIR communities and advocating data/digital object sharing has been the main strategic endeavour. FAIR by itself is not a goal but rather a process leading to open science and reproducible scientific practice. As the FAIR principles are relatively recently adopted in research, there is a transitional challenge in adapting and following them. This is mainly due to the decentralized definitions of what constitutes FAIR for AI models and other digital objects [12]. Circling around the four principles of FAIR, different suggestions were made by researchers [1,24,25].

However, in medical and epidemiological domains, following these suggestions become less practical. For example, in order to publish the result of an observational study conducted on predicting factors of “x” disease using “y” algorithm on a population of “z”, researchers should follow the STORBE (Strengthening the Reporting of Observational studies in Epidemiology) reporting guideline [26], which structures reporting the important elements of the study. Thus, the reviewers and academic editors have a common stance whether the study followed the appropriate methodology and reported the results based on the predefined expectations from observational studies. To accommodate the FAIR sharing of models and data, the authors have to go beyond the journals’ predefined expected requirements which usually is ignored and leads to potentially

irreproducible results. This is not unique for observational studies but also clinical trials and other experimental studies that involves AI interventions.

To mitigate this, we suggest calibration of existing reporting guidelines with FAIR principles. Here, we introduced a calibration framework consisting of a structured flow for mapping reporting guideline to FAIR principles. This mapping facilitates a transparent alignment between the guidelines' recommendations and the FAIR principles. In this way, we can integrate FAIR sharing practice in research methodologies that involve AI interventions and harness the benefits of open science in the long run [27]. The argument here is that instead of developing additional reporting guidelines, we should tune the already available ones to adapt the recent changes in the field.

The current work achieves an important first milestone in describing the core steps in calibrating guidelines with FAIR principles. It is part of an ongoing research effort aiming to integrate FAIR principles in reporting guidelines. We further plan to expand upon these initial findings and implement the calibration framework for several reporting guidelines. Through these efforts, we believe that the calibrated guidelines contribute to an improved culture towards open and reproducible science.

Conclusions

Our work lay the foundation for a novel approach to advancing reproducibility in AI research. By integrating FAIR principles with established reporting guidelines, the proposed tuning frame bridges the gap in accommodating both FAIR metrics and reporting frameworks and benefits from advantages of both major integrated components.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

References

- Samuel, S., F. Löffler, and B. König-Ries. *Machine Learning Pipelines: Provenance, Reproducibility and FAIR Data Principles*. 2021. Cham: Springer International Publishing.
- Hutson, M., *Artificial intelligence faces reproducibility crisis*. 2018, American Association for the Advancement of Science.
- Levinson, M.A., et al., *FAIRSCAPE: a Framework for FAIR and Reproducible Biomedical Analytics*. *Neuroinformatics*, 2022. **20**(1): p. 187-202.
- Wagner, A.S., et al., *FAIRly big: A framework for computationally reproducible processing of large-scale data*. *Scientific Data*, 2022. **9**(1): p. 80.
- Kapoor, S. and A. Narayanan, *Leakage and the reproducibility crisis in ML-based science*. arXiv preprint arXiv:2207.07048, 2022.
- Baker, M., *Reproducibility crisis*. *Nature*, 2016. **533**(26): p. 353-66.
- Thibeau-Sutre, E., et al., *ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing*. *Computer Methods and Programs in Biomedicine*, 2022. **220**: p. 106818.
- Hutson, M., *Artificial intelligence faces reproducibility crisis*. *Science*, 2018. **359**(6377): p. 725-726.
- Shelmerdine, S.C., et al., *Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare*. *BMJ Health & Care Informatics*, 2021. **28**(1).
- Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship*. *Scientific Data*, 2016. **3**(1): p. 160018.
- Ravi, N., et al., *FAIR principles for AI models with a practical application for accelerated high energy diffraction microscopy*. *Scientific Data*, 2022. **9**(1): p. 657.
- Huerta, E., et al., *FAIR for AI: An interdisciplinary and international community building perspective*. *Scientific Data*, 2023. **10**(1): p. 487.
- Bahim, C., et al., *The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments*. 2020.
- Carroll, C., A. Booth, and K. Cooper, *A worked example of "best fit" framework synthesis: a systematic review of views concerning the taking of some potential chemopreventive agents*. *BMC medical research methodology*, 2011. **11**(1): p. 1-9.
- Network, E., *Reporting guidelines*. United Kingdom, 2019.
- Shiferaw, K.B., et al., *Guidelines and Standard Frameworks for AI in Medicine: Protocol for a Systematic Literature Review*. *JMIR Res Protoc*, 2023. **12**: p. e47105.
- Wang, Y., et al., *Guidelines, Consensus Statements, and Standards for the Use of Artificial Intelligence in Medicine: Systematic Review*. *J Med Internet Res*, 2023. **25**: p. e46089.

18. Brouwers, M.C., et al., *AGREE II: advancing guideline development, reporting and evaluation in health care*. Cmaj, 2010. **182**(18): p. E839-E842.
19. Page, M.J., et al., *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews*. International journal of surgery, 2021. **88**: p. 105906.
20. Liu, X., et al., *Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension*. The Lancet Digital Health, 2020. **2**(10): p. e537-e548.
21. RDA FAIR Data Maturity Model Working Group, B., *FAIR Data Maturity Model: specification and guidelines*. Res. Data Alliance, 2020. **10**.
22. Carroll, C., et al., *"Best fit" framework synthesis: refining the method*. BMC medical research methodology, 2013. **13**(1): p. 1-16.
23. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific data, 2016. **3**(1): p. 1-9.
24. Goble, C., et al., *FAIR Computational Workflows*. Data Intelligence, 2020. **2**(1-2): p. 108-121.
25. Hong, N.P.C., et al., *FAIR principles for research software (FAIR4RS principles)*. 2022.
26. Vandembroucke, J.P., et al., *Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration*. Annals of internal medicine, 2007. **147**(8): p. W-163-W-194.
27. Dempsey, W., et al., *Sharing begins at home: How continuous and ubiquitous FAIRness can enhance research productivity and data reuse*. Harvard data science review, 2022. **4**(3).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.