JAMIA Open, 2024, **7(4)**, ooae105
https://doi.org/10.1093/jamiaopen/ooae105
Brief Communications

Brief Communications

CALIFRAME: a proposed method of calibrating reporting guidelines with FAIR principles to foster reproducibility of AI research in medicine

Kirubel Biruk Shiferaw, MPH[1,*], Irina Balaur, PhD[2], Danielle Welter, PhD[3], Dagmar Waltemath, PhD[1], Atinkut Alamirrew Zeleke, PhD[1]

[1]Department of Medical Informatics, Institute for Community Medicine, University Medicine Greifswald, Greifswald D-17475, Germany, [2]Luxembourg Centre for Systems Biology, University of Luxembourg, Belvaux L-4367, Luxembourg, [3]Luxembourg National Data Service, Esch-sur-Alzette L-4362, Luxembourg

*Corresponding author: Kirubel Biruk Shiferaw, MPH, Department of Medical Informatics, University Medicine Greifswald, Walther-Rathenau-Str. 48, Greifswald D 17475, Germany (s-kishif@uni-greifswald.de)

Abstract

Background: Procedural and reporting guidelines are crucial in framing scientific practices and communications among researchers and the broader community. These guidelines aim to ensure transparency, reproducibility, and reliability in scientific research. Despite several methodological frameworks proposed by various initiatives to foster reproducibility, challenges such as data leakage and reproducibility remain prevalent. Recent studies have highlighted the transformative potential of incorporating the FAIR (Findable, Accessible, Interoperable, and Reusable) principles into workflows, particularly in contexts like software and machine learning model development, to promote open science.

Objective: This study aims to introduce a comprehensive framework, designed to calibrate existing reporting guidelines against the FAIR principles. The goal is to enhance reproducibility and promote open science by integrating these principles into the scientific reporting process.

Methods: We employed the "Best fit" framework synthesis approach which involves systematically reviewing and synthesizing existing frameworks and guidelines to identify best practices and gaps. We then proposed a series of defined workflows to align reporting guidelines with FAIR principles. A use case was developed to demonstrate the practical application of the framework.

Results: The integration of FAIR principles with established reporting guidelines through the framework effectively bridges the gap between FAIR metrics and traditional reporting standards. The framework provides a structured approach to enhance the findability, accessibility, interoperability, and reusability of scientific data and outputs. The use case demonstrated the practical benefits of the framework, showing improved data management and reporting practices.

Discussion: The framework addresses critical challenges in scientific research, such as data leakage and reproducibility issues. By embedding FAIR principles into reporting guidelines, the framework ensures that scientific outputs are more transparent, reliable, and reusable. This integration not only benefits researchers by improving data management practices but also enhances the overall scientific process by promoting open science and collaboration.

Conclusion: The proposed framework successfully combines FAIR principles with reporting guidelines, offering a robust solution to enhance reproducibility and open science. This framework can be applied across various contexts, including software and machine learning model development stages, to foster a more transparent and collaborative scientific environment.

Lay Summary

This brief communication addresses the need for clear and trustworthy reporting in Artificial Intelligence (AI) research, particularly in healthcare. As AI continues to evolve, ensuring that research findings are transparent and reproducible is important. Current reporting practices often fall short, leading to overly optimistic claims about AI models. To tackle this issue, we propose a new framework that aligns existing reporting guidelines with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. Our proposed framework aims to enhance the clarity and reliability of AI research by integrating these principles into the reporting process without changing the core content of the reporting guidelines. We developed this framework through literature review and iterative feedback from experts. The result is a set of defined steps for calibrating reporting guidelines, demonstrated through a use case. By harmonizing reporting standards with FAIR principles, our framework not only improves the management and sharing of research data but also fosters a culture of collaboration and shared knowledge across various fields.

Key words: 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"

Received: August 8, 2024; **Revised:** September 20, 2024; **Editorial Decision:** September 25, 2024; **Accepted:** September 26, 2024

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 workflow to develop the calibrated reporting guideline is represented in Figure 1.

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

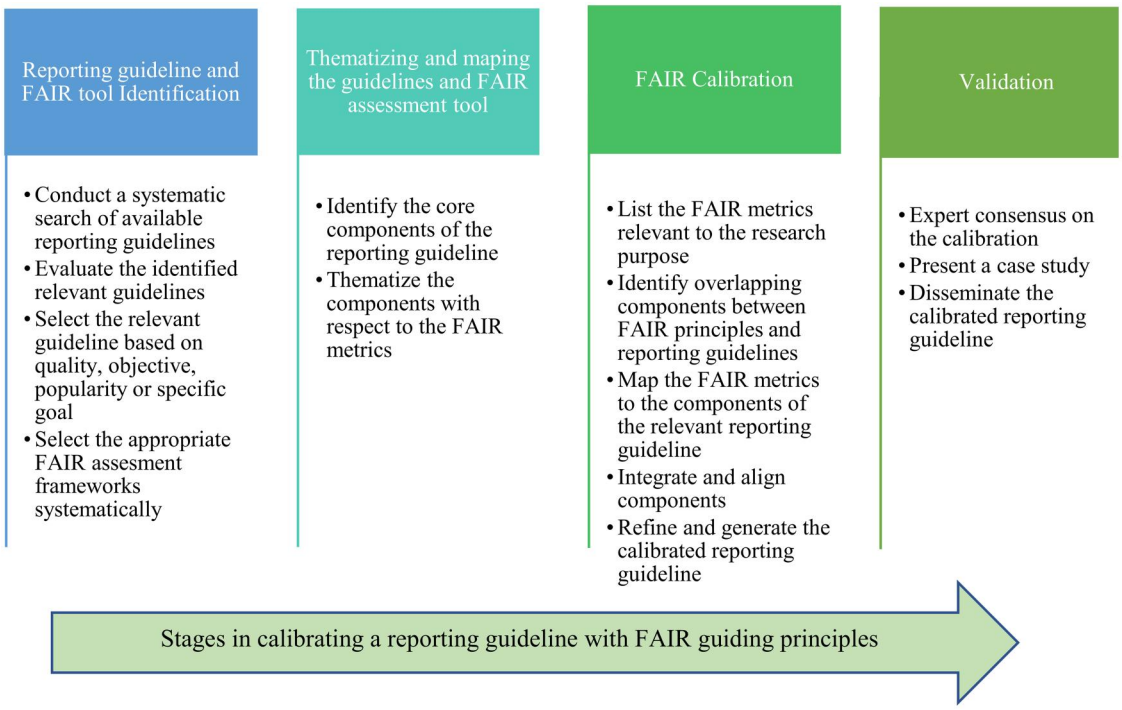

Figure 1. Diagrammatic representation of the stages for calibrating the existing reporting guidelines with the FAIR principles.

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.[20] For demonstration purposes, we selected the Consolidated Standards of Reporting Trials-Artificial Intelligence extension (CONSORT-AI)[21] guideline and the Research Data Alliance (RDA) FAIR Data Maturity Model.[22] 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.[22] 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.[22]

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.[23] To do so, a profound understanding of the FAIR metrics and the identified guideline is required. To ensure a transparent and robust calibration, we recommend a series of workshops involving a diverse group of experts in guidelines and specialists in FAIR principles within the context of machine learning and research software. These workshops will provid a collaborative environment for in-depth discussions and evaluations of the various guidelines.

Furthermore, these workshops facilitate continuous feedback, enabling the expert group to refine the calibration output after each session. Throughout the process, a meticulous documentation of the discussions, decisions made, and the rationale for including or excluding specific components of the guideline is highly recommended. This documentation serves as a transparent record of the calibration process, allowing users to trace the provenance back to the original guidelines. Finally, a consensus-building review session is required to validate the findings, ensuring that the final recommendations reflect a collective agreement among the experts.

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). These indicators emphasize the importance of making data and metadata easily discoverable, which is crucial for fostering reproducibility and trust. Based on this evaluation, we suggest a solution to calibrate the integration of these frameworks, ensuring that the strengths of both the RDA FAIR indicators and CONSORT-AI items are leveraged effectively. The possible calibration method could be rephrasing items to incorporate elements of FAIR principles or adding additional subitems. For instance, Item CONSORT-AI 25 Extension ("State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use") could be more enriched by adding sub items and detailed elaboration document specific to datasets and code (example: 25a: Information about AI intervention source code access condition (restricted access, open access, closed access), 25b: Access protocol information that describes the actions to be taken (if not open access), 25b1: Does the access protocol supports authentication and authorisation?). It is also important to note that some items might not align or map to any FAIR indicators. For example, items under randomization section in CONSORT AI (such as 6a-12b) does not align with any of the FAIR principles which means they are unique items for the reporting guideline. In this case the item should be kept for the next stage of the calibration process.

Such an evaluation not only provides valuable input for developing a calibrated integrated framework but also contributes to the ongoing discourse on best practices in AI research and data management.

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.

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.

Expected challenges

Calibrating reporting guidelines with FAIR principles presents several expected challenges that must be addressed to ensure a successful integration. The first one is the complexity of aligning these frameworks, as some items may not have clear counterparts, complicating the integration effort. Variability in interpretation among stakeholders could also lead to inconsistencies in application. Furthermore, the calibration might increase the number of items in the original reporting guideline which might affect it usability as researchers and potential users of the reporting guidelines prefer a comprehensive and likely shorter list of items. Thus, the calibration effort should be carefully managed to strike a balance between comprehensiveness and usability.

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.[10] 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.[12,24] Mobilizing FAIR communities and advocating data/digital object sharing has been the main strategic endeavor. 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.[1,24–26]

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,[27] 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 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.[28] 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.

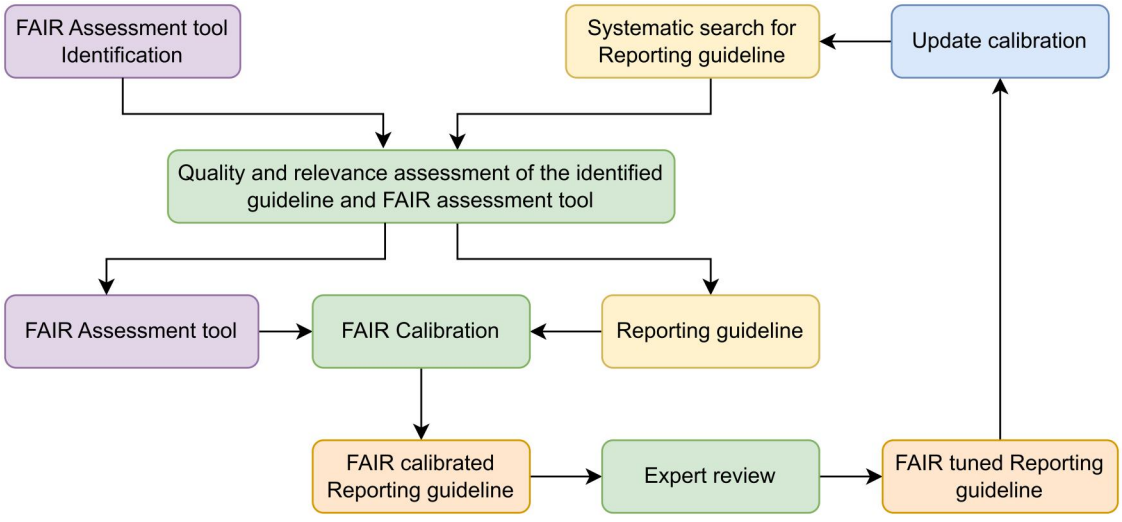

Figure 2. The FAIR calibration process of reporting guidelines. Identification of the FAIR assessment tool and reporting guideline can be performed in parallel. Similarly, the quality and relevance assessment of the identified guideline and FAIR assessment tool also can be done in parallel or one after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

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 toward open and reproducible science.

Conclusion

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.

Acknowledgments

K.B.S. would like to acknowledge the German Academic Exchange service (DAAD) for funding the research expenses.

Author contributions

Kirubel Biruk Shiferaw, Dagmar Waltemath, and Atinkut Alamirrew Zeleke contributed substantially to the conception, methodology and writing of this work. Irina Balaur and Danielle Welter, contributed significantly in revising, editing and writing the manuscript. All authors revised this manuscript critically for important intellectual content, approved the version to be published.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest

None declared.

Data availability

No new data were generated or analyzed in support of this research.

References

1. Samuel S, Löffler F, König-Ries B. *Machine Learning Pipelines: Provenance, Reproducibility and FAIR Data Principles*. Springer International Publishing; 2021.
2. Hutson M. *Artificial Intelligence Faces Reproducibility Crisis*. American Association for the Advancement of Science; 2018.
3. Levinson MA, Niestroy J, Al Manir S, et al. FAIRSCAPE: a framework for FAIR and reproducible biomedical analytics. *Neuroinformatics*. 2022;20:187-202.
4. Wagner AS, Waite LK, Wierzba M, et al. FAIRly big: a framework for computationally reproducible processing of large-scale data. *Sci Data*. 2022;9:80.
5. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in ML-based science, arXiv, arXiv:2207.07048, preprint: not peer reviewed; 2022.
6. Baker M. Reproducibility crisis. *Nature*. 2016;533:353-366.
7. Thibeau-Sutre E, Díaz M, Hassanaly R, et al. ClinicaDL: an open-source deep learning software for reproducible neuroimaging processing. *Comput Methods Programs Biomed*. 2022;220:106818.
8. Hutson M. Artificial intelligence faces reproducibility crisis. *Science*. 2018;359:725-726.
9. Shelmerdine SC, Arthurs OJ, Denniston A, et al. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. *BMJ Health Care Inform*. 2021;28:e100385.
10. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
11. Ravi N, Chaturvedi P, Huerta EA, et al. FAIR principles for AI models with a practical application for accelerated high energy diffraction microscopy. *Sci Data*. 2022;9:657.
12. Huerta EA, Blaiszik B, Brinson LC, et al. FAIR for AI: an interdisciplinary and international community building perspective. *Sci Data*. 2023;10:487.
13. Bahim C, Casorrán-Amilburu C, Dekkers M, et al. The FAIR data maturity model: an approach to harmonise FAIR assessments. *Data Sci J*. 2020;19:41.
14. Carroll C, Booth A, Cooper K. A worked example of" best fit" framework synthesis: a systematic review of views concerning the taking of some potential chemopreventive agents. *BMC Med Res Methodol*. 2011;11:29.
15. The Equator Network. "What Is a Reporting Guideline?" Equator network. Accessed October 10, 2024. https://www.equator-network.org/about-us/what-is-a-reporting-guideline/
16. Shiferaw KB, Roloff M, Waltemath D, et al. Guidelines and standard frameworks for AI in medicine: Protocol for a systematic literature review. *JMIR Res Protoc*. 2023;12:e47105.
17. Wang Y, Li N, Chen L, et al. Guidelines, consensus statements, and standards for the use of artificial intelligence in medicine: Systematic review. *J Med Internet Res*. 2023;25:e46089.
18. Brouwers MC, Kho ME, Browman GP, et al.; AGREE Next Steps Consortium. AGREE II: advancing guideline development, reporting and evaluation in health care. *CMAJ*. 2010;182:E839-E842.
19. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int J Surg*. 2021;88:105906.
20. Shiferaw KB, Roloff M, Balaur I, et al. Guidelines and standard frameworks for artificial intelligence in medicine: a systematic review, medRxiv, https://doi.org/10.1101/2024.05.27.24307991, preprint: not peer reviewed; 2024.
21. Liu X, Cruz Rivera S, Moher D, et al.; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digital Health*. 2020;2:e537-e548.
22. RDA FAIR Data Maturity Model Working Group. FAIR data maturity model: specification and guidelines. *Res Data Alliance*. 2020;10. https://doi.org/10.15497/RDA00050
23. Carroll C, Booth A, Leaviss J, et al. "Best fit" framework synthesis: refining the method. *BMC Med Res Methodol*. 2013;13:1-16.
24. Ammar A, Evelo C, Willighagen E. FAIR assessment of nanosafety data reusability with community standards. *Sci Data*. 2024;11:503.
25. Goble C, Cohen-Boulakia S, Soiland-Reyes S, et al. FAIR computational workflows. *Data Intell*. 2020;2:108-121.
26. Barker M, Chue Hong NP, Katz DS, et al. Introducing the FAIR Principles for research software. *Sci Data*. 2022;9:622. https://doi.org/10.1038/s41597-022-01710-x
27. Vandenbroucke JP, von Elm E, Altman DG, et al.; STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Ann Int Med*. 2007;147:W163-W194.
28. Dempsey WP, Foster I, Fraser S, Kesselman C. Sharing begins at home: How continuous and ubiquitous FAIRness can enhance research productivity and data reuse. *Harv Data Sci Rev*. 2022;4. https://doi.org/10.1162/99608f92.44d21b86

JAMIA Open, 2024, 7, 1–5
https://doi.org/10.1093/jamiaopen/ooae105
Brief Communications