Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230065

Assessing the FAIRness of Deep Learning Models in Cardiovascular Disease Using Computed Tomography Images: Data and Code Perspective

Kirubel Biruk SHIFERAW[a,1], Atinkut ZELEKE[a] and Dagmar WALTEMATH[a]
[a] *Medical Informatics Laboratory, Institute for Community Medicine, University Medicine Greifswald, Germany*

ORCiD ID: Kirubel Biruk Shiferaw https://orcid.org/0000-0002-7411-1411
Atinkut Zeleke https://orcid.org/0000-0001-7838-9050
Dagmar Waltemath https://orcid.org/0000-0002-5886-5563

Abstract. The interest in the application of AI in medicine has intensely increased over the past decade with most of the changes in the past five years. Most recently, the application of deep learning algorithms in prediction and classification of cardiovascular diseases (CVD) using computed tomography (CT) images showed promising results. The notable and exciting advancement in this area of study is, however, associated with different challenges related to the findability (F), accessibility(A), interoperability(I), reusability(R) of both data and source code. The aim of this work is to identify reoccurring missing FAIR-related features and to assess the level of FAIRness of data and models used to predict/diagnose cardiovascular diseases from CT images. We evaluated the FAIRness of data and models in published studies using the RDA (Research Data Alliance) FAIR Data maturity model and FAIRshake toolkit. The finding showed that although AI is anticipated to bring ground breaking solutions for complex medical problems, the findability, accessibility, interoperability and reusability of data/metadata/code is still a prominent challenge.

Keywords. FAIR Principles, Deep learning, cardiovascular disease, computed tomography, RDA FAIR Data maturity model

1. Introduction

The development of computational models using Artificial intelligence (AI) in Medicine has gained high interest in the last five years due to the new possibilities to incorporate multi-modal biomedical data as well as to mimic and to explore the complexity of the events and interdependencies at various levels (molecular, cellular, tissue/ organ, whole-body) of the human biomedical systems [1]. This development has opened new paths in approaching medical problems with complex and robust AI

[1] Corresponding Author: Kirubel Biruk Shiferaw, Walther-Rathenau-Str. 48, Medical Informatics Laboratory, University Medicine Greifswald, D-17475 Germany; E-mail: s-kishif@uni-greifswald.de.

applications in terms of virtual (model) and physical (device) methods. Deep learning (DL) revolutionized the application of AI in medicine, especially in image processing [2].

Most recently, the application of DL algorithms in cardiovascular disease (CVD) risk/event predication and classification using SPECT/CT (A single-photon Emission Computed Tomography) and PET/CT (Positron Emission Tomography) images showed promising results [3]. Given the advancement in scanner technology in both image quality and dimensions, Computed Tomography (CT) is well suited for advanced image analysis using deep neural networks [4]. Indeed, the published research outputs show an increase of approaches for predicting and diagnosing CVD using CT imaging.

However, the reproducibility of DL-based studies has become a challenge [5]. It is consistently mentioned in the literature that reproducibility is a core issue in the scientific process, and this includes AI research [6]. The notable and exciting advancement in this area of study is associated with different challenges relating to the findability (F), accessibility(A), interoperability(I), reusability(R) of both data and source code.

The FAIR guiding principles are one of the recently applied set of guidelines to facilitate the discovery and reuse of scientific digital objects including data, metadata, software and tools [7]. Substantial effort by different working initiatives has been made to quantify FAIRness of digital objects [8], and the FAIR evaluation tools have been developed to help identify weak points in data and code representation within particular scientific domains (e.g. https://fairassist.org/#!/) [9]. Beside contributing to the reproducibility of published studies, the sharing of data and code facilitates rigor scientific practice and reassures the validity of the claimed results [10]. Using standard frameworks, adhering to guiding principles, and developing and reporting guidelines are some of the most common approaches of standardizing data and code sharing, e.g. as demonstrated by the "Computational Modeling in Biology" Network (COMBINE: https://co.mbine.org/) community in the area of computational biology [11, 12].

We aim to assess the level of FAIRness of current DL models in imaging in CVD, and to identify systemic lacks of the FAIR principles which might lead us to recommend coordinated actions for future research in the field. Therefore, we used the RDA (Research Data Alliance) FAIR Data maturity model [13] and FAIRshake tools [14] to evaluate the FAIRness of DL models and associated data in studies on CVDs from CT images.

2. Method

First, we defined the following set of keywords to describe the CVD-applied DL models of interest: Cardiovascular, CT scan, Deep learning, Diagnosis, Heart Defect, Computed, tomography, Hierarchical learning, classification, Congenital, X-ray computed, prognosis and prognosis. After keywords had been identified, using this keyword list, we performed a comprehensive search in Web of Science within the time frame of 2016-2022. We included studies conducted with DL models to predict/classify cardiovascular disease/event from PET(CT)/SPCT(CT) images. As such, a list of 109 publications with their corresponding models was used for further descriptive analysis and FAIRness assessment using essential indicators of the RDA FAIR Data maturity model and the FAIRshake tool. The RDA FAIR Data maturity model is an evaluation tool to assess adherence to the FAIR principles. The available indicators provide three different

degrees with respect to impact on the FAIRness, namely, Essential (critical to achieve FAIRness), Important (with high contribution to the FAIRness features), or Useful (increase the overall FAIRness level of the resources) [15]. The RDA FAIR Data model furthermore offers a scale-based approach to prioritize and self-evaluate the level of FAIRness. FAIRshake is also another tool developed to facilitate the establishment of community driven metrics and rubrics paired with manual and automated FAIR assessment with insignia visualization [14].

3. Results

The search resulted in 109 articles. After excluding studies with closed access, in non-English language, and studies that were irrelevant for our objective, only 22 studies were further analyzed. With respect to datasets associated with the respective studies, only 5/22 (22.7%) provide a URL for the data; 7/22 (31.8%) provide neither metadata nor URL at all in the document. Furthermore, only 4/22 (18.2%) studies provide information on how to access the data used in the study such as "Available on reasonable request from the Author". Finally, 4/22 (18.2%) studies provided both URL and metadata. With respect to code availability, only 2/22 (9.1%) studies made their code available for reuse with license details and only one study stated that the code is "available on reasonable request". Only 2/22 (9.1%) of the studies used a reporting standard namely TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) [16] and STARD (The Standards for Reporting of Diagnostic Accuracy) [17].

The analysis from the RDA-FAIR Data maturity model (Figure 1) shows that most of the essential indicators were not satisfied, particularly metadata related indicators were poorly represented.

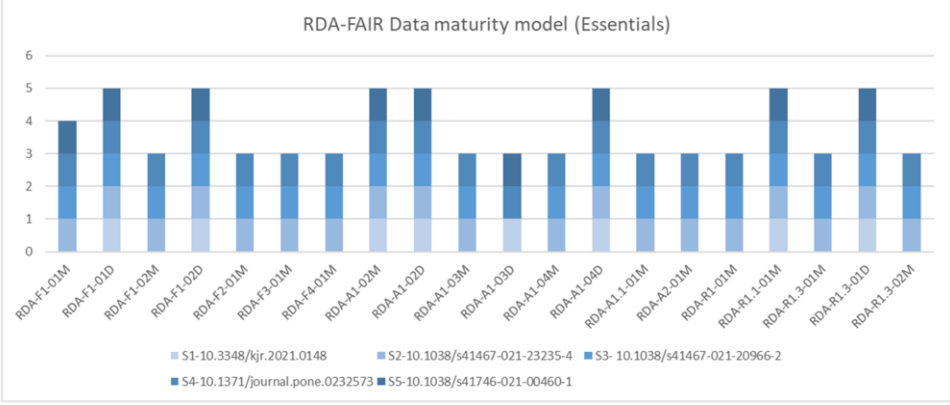

Figure 1. RDA-FAIR data maturity model assessment (x-axis: Essential indicators, y-axis: number of studies satisfying specific indicators)

The FAIRshake insignia visualization (Figure 2) is based on FAIR metrics by FAIRmetrics.org. It also shows that a satisfactory level of FAIRness is still not reached in the scientific field of DL studies on CVD using CT image. Most squares in Figure 2 are red (hence do not satisfy the FAIR metrics) with most of the challenges in Findability and Reusability sectors.

Figure 2. FAIRshake insignia assessment. Each set of four squares of the insignia from the left upper corner to right lower corner represents one aspect of FAIR [(F, A, I&R)].

4. Discussion

The advancement in AI applications, especially DL algorithms in the medical and healthcare domains, is becoming a promising asset for precise and personalized care [18]. However, our first assessment of the *status quo* shows that the studies evaluated on this work do not share their data or code/software; the studies also do not provide protocols. It would be worth conducting further research to determine the extent to which the lack of code and data sharing is a general trend, rather than specific to the domain of CVD. It is understandable that patient data is sensitive and the developed algorithms are intellectual properties [19]. However, sharing does not necessarily mean providing unlimited access for free, rather using a set of protocols and appropriate licenses that enable other researchers to use and cite the work as needed.

We noticed that some researchers tend to share a URL for their data and code which is a good starting point. However, a dataset/code without detailed metadata is not easily reusable as context information is missing. We argue that researchers should publish a detailed standardized metadata along with their research outcomes to facilitate FAIRness of their resources and to potentially increase the reproducibility [20] and reusability of DL models in CVD. It is also important to note that if results and models are irreproducible, it is very likely that efforts will need to be duplicated, resulting in extra monetary costs, longer time to publication and less trust [21, 22].

It is important to know that a FAIR assessment does not assess the quality of methodology or result of the studies, but it indicates how well a study adheres to modern research data management practices in terms of findability, accessibility, interoperability and reusability of the associated data/code. While a detailed comparison of the functionality of the two methods is beyond the scope of this work, the similar results from both approaches suggest that further research on this topic could be insightful

We therefore recommend biomedical scientists to use available FAIR assessment tools, including the RDA FAIR data maturity model, to evaluate their own works, preferably before publication. Help in using these tools and deriving a data management strategy can be obtained from data stewards at research institutions, data integration centers, or community work groups.

5. Conclusion

Although AI is anticipated to bring ground breaking solutions for complex medical problems, adherence to the FAIR principles for data/metadata/code is still a prominent challenge. Authors should consider standardized ways of sharing their data/metadata/code. Other stakeholders in the publishing ecosystem such as reviewers, editors and publishers should encourage FAIR sharing for the interest of reproducibility, trust and ultimately for the advancement of open science.

References

[1] Shiferaw, K.B., D. Waltemath, and A. Zeleke, Disparities in Regional Publication Trends on the Topic of Artificial Intelligence in Biomedical Science Over the Last Five Years: A Bibliometric Analysis. Studies in Health Technology and Informatics, 2022. **294**: p. 609-613.

[2] Fu, Y., et al., A review of deep learning based methods for medical image multi-organ segmentation. Physica Medica, 2021. **85**: p. 107-122.

[3] Slart, R.H., et al., Position paper of the EACVI and EANM on artificial intelligence applications in multimodality cardiovascular imaging using SPECT/CT, PET/CT, and cardiac CT. European journal of nuclear medicine and molecular imaging, 2021. **48**(5): p. 1399-1413.

[4] Lin, A., et al., Artificial intelligence in cardiovascular CT: Current status and future implications. Journal of cardiovascular computed tomography, 2021. **15**(6): p. 462-469.

[5] Hartley, M. and T.S. Olsson, dtoolai: Reproducibility for deep learning. Patterns, 2020. **1**(5): p. 100073.

[6] Gundersen, O.E. and S. Kjensmo. State of the art: Reproducibility in artificial intelligence. in Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[7] Wilkinson, M.D., et al., The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 2016. **3**(1): p. 1-9.

[8] Wilkinson, M.D., et al., Evaluating FAIR maturity through a scalable, automated, community-governed framework. Scientific data, 2019. **6**(1): p. 1-12.

[9] Sansone, S.-A., et al., FAIRsharing as a community approach to standards, repositories and policies. Nature Biotechnology, 2019. **37**(4): p. 358-367.

[10] Peng, R.D., F. Dominici, and S.L. Zeger, Reproducible epidemiologic research. American journal of epidemiology, 2006. **163**(9): p. 783-789.

[11] Waltemath, D., et al., The first 10 years of the international coordination network for standards in systems and synthetic biology (COMBINE). Journal of integrative bioinformatics, 2020. **17**(2-3).

[12] Niarakis, A., et al., Addressing barriers in comprehensiveness, accessibility, reusability, interoperability and reproducibility of computational models in systems biology. Briefings in Bioinformatics, 2022.

[13] Alliance, R.D. RDA FAIR Data maturity model. [cited 2022; Available from: https://www.rd-alliance.org/groups/fair-data-maturity-model-wg.

[14] Clarke, D.J., et al., FAIRshake: toolkit to evaluate the FAIRness of research digital resources. Cell systems, 2019. **9**(5): p. 417-421.

[15] FAIR Data Maturity Model Working Group, FAIR Data Maturity Model. Specification and Guidelines (1.0), 2020. https://doi.org/10.15497/rda00050.

[16] Collins, G.S., et al., Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Journal of British Surgery, 2015. **102**(3): p. 148-158.

[17] Bossuyt, P.M., et al., Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Annals of internal medicine, 2003. **138**(1): p. 40-44.

[18] Razzak, M.I., S. Naz, and A. Zaib, Deep learning for medical image processing: Overview, challenges and the future. Classification in BioApps, 2018: p. 323-350.

[19] Jean-Paul, S., et al. Issues in the Reproducibility of Deep Learning Results. in 2019 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). 2019.

[20] Hartley, M. and T.S.G. Olsson, dtoolAI: Reproducibility for Deep Learning. Patterns, 2020. **1**(5): p. 100073.

[21] Gundersen, O.E., The Reproducibility Crisis Is Real. AI Mag., 2020. **41**(3): p. 103-106.

[22] König, M., et al. Challenges and opportunities for system biology standards and tools in medical research. in 7th Workshop on Ontologies and Data in Life Sciences, ODLS 2016. 2016. CEUR-WS.