

The T.R.U.E. Checklist for Identifying Impactful Artificial Intelligence–Based Findings in Nuclear Medicine: Is It True? Is It Reproducible? Is It Useful? Is It Explainable?

Irène Buvat and Fanny Orlhac

Institut Curie, Université PSL, Inserm U1288, Laboratory of Translational Imaging in Oncology, Orsay, France

Dozens of articles describing artificial intelligence (AI) developments are submitted to medical imaging journals every month, including in the nuclear medicine field. Our mission, as a nuclear medicine community, is to contribute to a better understanding of normal and pathologic processes by probing molecular mechanisms with an unparalleled sensitivity, ultimately with the goal of improving patient care. This mission calls for research in tracer development, instrumentation, data analysis, and clinical studies. It is becoming obvious that our mission will be greatly facilitated by AI-based tools. It is far too early to estimate the exact impact AI will have on nuclear medicine research and clinical practice. Still, we can already claim that AI will assist in the automation of many tasks, including image acquisition, image interpretation, and image quantification, hence increasing the reproducibility, overall quality, and usefulness of nuclear medicine scans (1–3). Less clear is whether AI can be used to further biomedical knowledge, such as through better understanding of molecular mechanisms or identification of new clinically useful biomarkers involving nuclear medicine data. So far, in nuclear medicine, no new biomarkers involving sophisticated radiomic features or deep learning models have emerged from the thousands of articles already published. None of the published promising radiomic signatures, nomograms, or AI-based models have been convincingly demonstrated by independent groups as must-have biomarkers superior to existing practice based on large-scale evaluation. Yet, we trust that this goal is within reach. AI has demonstrated its ability to identify and reveal complex information hidden in images, and it should be possible to use this information to extract clinically useful biomarkers. To get to this point, we have to be extremely demanding in terms of what is published so that the most promising findings can easily be identified by readers. This ability would allow the community to subsequently gather the large body of evidence needed to turn a promising result into an actionable biomarker, a testable assumption, or a widely used automated method. To facilitate the identification of those contributions that might be ground-breaking, we encourage the authors and reviewers of AI-based manuscripts to carefully consider a simple checklist—the T.R.U.E. checklist—in which the acronym T.R.U.E. comprises 4

questions: Is it True? Is it Reproducible? Is it Useful? Is it Explainable? A “yes” answer to all 4 questions increases the likelihood that the reporting will be impactful. In fact, these 4 questions should be part of every professional review process of any scientific paper—whatever the research topic—and have been extensively used in the past. Yet, they are of particular and critical relevance to papers using AI-based methods, because of the specifics of AI. We now briefly elaborate on these questions to more precisely explain what they imply in the context of AI-based studies.

IS IT TRUE?

The question of truth is highly relevant because a large proportion of AI-based studies in medical imaging are still biased by issues well known to data scientists, such as bias in the training population (e.g., sex, ethnicity, and age), data leakage (i.e., test data used explicitly or implicitly during the training phase) (4), or overfitting. This bias most often results in a lack of generalizability of the AI-based model, meaning that the results and reported level of performance will not hold on different datasets (5). By default, we should assume that the findings, especially when outstanding, are biased, and we should chase potential confounding factors by all means. Control experiments (similar to experiments using a sham group or placebo arm in clinical trials) should be used and reported whenever relevant, giving enough evidence that the findings are scientifically valid. For instance, the probability of false-positive findings can be determined by repeating the extensive model-building and model-evaluation process after randomly permuting the label associated with each patient. Expert data scientists should be called on to assist in the possible identification of bias or sources of data leakage, given that these can be subtle and difficult to detect. Medical experts of course remain essential to detect bias or possible confounding effects associated with the composition of the patient samples.

IS IT REPRODUCIBLE?

The reproducibility crisis affects many fields and has been extensively studied and debated (6), including in the field of radiology (7,8). There have been laudable efforts over the last few years to increase transparency, with the very positive trend of data and models being shared more frequently, resulting in an overall improvement in radiomic and AI-based imaging study quality (9). Yet, even when authors share their models developed within well-known frameworks (e.g., TensorFlow or CAFFE [Convolutional

Received Dec. 8, 2020; revision accepted Mar. 9, 2021

For correspondence or reprints, contact Irène Buvat (irene.buvat@u-psud.fr).

Published online Mar. 26, 2021.

COPYRIGHT © 2021 by the Society of Nuclear Medicine and Molecular Imaging.

DOI: 10.2967/jnumed.120.261586

Architecture for Fast Feature Embedding]) using one of the many resource-sharing platforms (e.g., GitHub, SourceForge, GitLab, GitKraken, or Bitbucket), this sharing is often not sufficient to actually reproduce the findings, even when the data are also provided. One of the reasons is that most AI-based models are complex and involve many steps and parameters, such as those relating to image preprocessing, data augmentation, and learning schemes, and these are usually not fully described, despite significantly impacting the results. In AI, “the devil is in the details,” as the saying goes. To overcome this reproducibility challenge and move the field forward, we strongly encourage authors to carefully describe their methods and provide data or code (either source code or executable code) that might be needed to reproduce the investigation or test the model on independent data. In addition, similar to the current practice of calling on statistical expertise to validate the statistical methodology used in scientific manuscripts, we recommend calling on specific expertise to practically check that the provided description or material makes it possible to reproduce the findings and test the models on external data. This extra workload on the reviewers would hugely increase the value of published AI-based contributions. We expect contributions that report reproducible methods to have a much greater impact than those that do not.

IS IT USEFUL?

The usefulness should be appreciated with respect to state-of-the-art knowledge and methods, and a comparison of results with previously published data is a good way to assess the usefulness of new findings. Such comparisons can be difficult when different methods are not assessed on the same dataset, because of many possible confounding factors. Sharing of datasets, which can then be used as benchmarks to compare different methods, as in medical imaging challenges (10), can facilitate fair comparison. In what respect the new findings are superior to existing, and often simpler, methods should always be demonstrated. Performance analysis should include metrics characterizing the robustness of the method with respect to potential perturbations (e.g., data of different quality) so as to properly assess the trade-off between complexity, accuracy, and robustness achieved by different models. Occam’s razor should remain the rule until well-supported evidence of the superiority of less intuitive and more complex models is obtained. Although AI is extremely powerful, its power should rather be used when conventional statistical approaches or signal-processing methods are insufficient. There can be different motivations for using an AI model: an AI-based method can save time while equaling human observer performance (11), it can equal human observer performance while reducing interobserver variability (12), it might outperform existing human-based performance (13) or algorithm-based performance (14) (although this will have to be proven in prospective studies), or it might even

NOTEWORTHY

- AI algorithms are currently proposed for many different purposes in nuclear medicine.
- The reporting of these algorithms poses special challenges that require appropriate transparency and a high level of scientific rigor.
- Any report involving an AI-based method should carefully address and discuss the scientific validity, reproducibility, usefulness, and explainability of the findings.

uncover unknown phenomena (15). Whatever the scenario, the added value of the AI model should be well substantiated.

IS IT EXPLAINABLE?

AI is not a magic wand. It is a powerful set of algorithms that learn from examples and have the unique ability to identify structures in highly dimensional data. When AI is used to automate a task that humans can do by learning from many examples, the rules are deduced by the AI from the examples. The performance to be expected will depend on the representativity of the learning ensemble compared with the cases that can be encountered in practice. A more challenging application is having the AI succeed in doing something that we, as humans, cannot do (yet). As an example, we are currently at a loss when predicting why certain patients will respond to immunotherapy whereas others will not. For these applications, investigating what makes an AI algorithm successful is essential to avoid misinterpretation and prevent overestimation of the power of AI. For example, a misinterpretation of an AI decision-making process was published in a highly respected journal (16), before a reanalysis of the data elegantly demonstrated the incorrect understanding of the initial results (17). This error emphasizes the need for scrutiny of the key elements explaining the performance of an AI-based model. By better understanding the AI model and which specific information it uses, we might also gain knowledge of the biologic mechanisms involved. For this explanation step, speculation is still currently the rule. To use AI as a datascope that will help us better understand the molecular mechanisms based on image content, we have to go from speculation to hypothesis formulation and then hypothesis testing using appropriate *in silico*, *in vitro*, or *in vivo* experimental design.

Explainable AI is currently an extremely active area of research, with the ongoing development of numerous methods for approaching explainability (18), although fully satisfactory explanations may not always be feasible because of the high complexity and dimensionality of the data (19). The “Is it explainable?” question is thus certainly the most difficult one to answer convincingly. Yet, it should not be avoided and should be addressed whenever possible so that AI can help us learn from the data.

CONCLUSION

It is our conviction that the articles for which all 4 T.R.U.E. questions are convincingly addressed have a much higher likelihood of yielding significant advances in our field compared with papers that do not meet this requirement. We thus encourage all investigators and authors to take the time to reflect on this easy-to-remember checklist before submitting to *The Journal of Nuclear Medicine*, to write out well-supported evidence of their responses to these questions, and to adjust their claims accordingly. We also invite all the devoted reviewers of *The Journal of Nuclear Medicine* to keep this checklist in mind when reviewing articles involving AI algorithms. In addition, to further assist investigators in the development of sound and reproducible AI-based research, the AI task force of the Society of Nuclear Medicine and Molecular Imaging will soon release consensus recommendations underlying the specifics associated with nuclear medicine applications.

DISCLOSURE

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments and David Wallis for carefully proofreading the manuscript.

REFERENCES

1. Sibille L, Seifert R, Avramovic N, et al. ^{18}F -FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. *Radiology*. 2020;294:445–452.
2. Betancur J, Commandeur F, Motlagh M, et al. Deep learning for prediction of obstructive disease from fast myocardial perfusion SPECT: a multicenter study. *JACC Cardiovasc Imaging*. 2018;11:1654–1663.
3. Ding Y, Sohn JH, Kawczynski MG, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using ^{18}F -FDG PET of the brain. *Radiology*. 2019;290:456–464.
4. Wen J, Thibeau-Sutre E, Diaz-Melo M, et al. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Med Image Anal*. 2020;63:101694.
5. Reuzé S, Orhac F, Chargari C, et al. Prediction of cervical cancer recurrence using textural features extracted from ^{18}F -FDG PET images acquired with different scanners. *Oncotarget*. 2017;8:43169–43179.
6. Wallach JD, Boyack KW, Ioannidis JPA. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015-2017. *PLoS Biol*. 2018;16:e2006930.
7. Wright BD, Vo N, Nolan J, et al. An analysis of key indicators of reproducibility in radiology. *Insights Imaging*. 2020;11:65.
8. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. *Nature*. 2020;586:E14–E16.
9. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. *Eur J Nucl Med Mol Imaging*. 2019;46:2656–2672.
10. Challenges. Grand Challenge website. <https://grand-challenge.org/challenges/>. Accessed April 26, 2021.
11. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*. 2019;1:e271–e297.
12. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
13. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*. 2019;25:954–961.
14. van Dijk LV, Van den Bosch L, Aljabar P, et al. Improving automatic delineation for head and neck organs at risk by deep learning contouring. *Radiother Oncol*. 2020;142:115–123.
15. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2:158–164.
16. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
17. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: the need for safeguards. *Radiother Oncol*. 2019;130:2–9.
18. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub; 2021. <https://christophm.github.io/interpretable-ml-book/>
19. Bathaee Y. The artificial intelligence black box and the failure of intent and causation. *Harv J Law Technol*. 2018;31:890–938.