Journal of Nuclear Medicine, published on March 26, 2021 as doi:10.2967/jnumed.120.261586

The T.R.U.E. checklist for identifying impactful AI-based findings in nuclearmedicine: is it True? Is it Reproducible? Is it Useful? Is it Explainable?

Authors:

Irène Buvat[1] (PhD) Fanny Orlhac[1]

(PhD)

*Corresponding author: Irène Buvat, PhD
Email: irene.buvat@u-psud.fr

[1] Laboratory of Translational Imaging in Oncology, U1288 Inserm, Institut Curie,Université Paris Saclay, Orsay, France.

Irène Buvat ORCID ID: 0000-0002-7053-6471

Running title: The T.R.U.E. checklist in AI

Word counts: 2173 words

Noteworthy
- Artificial intelligence (AI) algorithms are currently proposed for many different purposes in Nuclear Medicine. (Page 2)
- The reporting of these algorithms poses special challenges that require appropriate transparency and a high level of scientific rigor (Pages 2-3).
- Any report involving an AI-based method should carefully address and discuss the scientific validity, reproducibility, usefulness, and explainability of the findings (Pages 3 to 7).

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 pathological processes by probing molecular mechanisms with an unparalleled sensitivity, ultimately with the goal of improving patient care. This 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 in 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 eg (*1-3*). Less clear is whether AI can be used to further biomedical knowledge, for example 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 be easily identified by readers. This 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 asimple checklist (the T.R.U.E. checklist) composed of four questions: Is it True? Is it Reproducible? Is it Useful? Is it Explainable? A “Yes” answer to all of the four questions increases the likelihood that the reporting will be impactful. In fact, thesefour questions should be part of every professional review process of any scientific paper whatever the research topic and have been extensively utilized in the past.

Yet, they are of particular and critical relevance to papers using AI-based methods,due to 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? This question is highly relevant as a large proportion of AI-based studies inmedical imaging are still biased by issues well known to data scientists, such as bias in the training population (eg, gender, ethnic, age), data leakage (ie test data used explicitly or implicitly during the training phase) eg (*4*), or overfitting. This 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 eg (*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 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 remain of course 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 eg (*6*), including in the field of radiology eg (*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 and 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 (eg, Tensorflow, Caffe) using one of the many resource-sharing platforms (eg, GitHub, SourceForge, Gitlab, Gitkraken, Bitbucket), this is often not sufficient to actually reproduce the findings. This is true even when the data are alsoprovided. One of the reasons is that most AI-based models are complex and involve many steps and parameters relating to image preprocessing, data augmentation, learning schemes etc., and these are usually not fully described, despite significantly impacting the results. In AI, as the saying goes, “the devil is in the details”. To overcome this reproducibility challenge and move the field forward, we strongly encourage authors to carefully describe their methods and provide data and/or code(either source code or executable) 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 employed in scientific manuscripts, we recommend calling on specific expertise to practically check that the provided description and/or material makes it possible to reproduce the findings and test the models on external data. This extraworkload on the reviewers would hugely increase the value of published AI-based contributions. We expect contributions reporting 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 datais 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, due to manypossible confounding factors. Sharing datasets, which can then be used as “benchmarks” to compare different methods, as in Medical Imaging challenges[1], canfacilitate 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 (eg, 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 employed when conventional statistical approaches or signal processing methods are insufficient. There can be different motivations for using an AI

[1] https://grand-challenge.org/challenges/

model: an AI-based method can save time whileequaling human observer performance eg (*10*), it can equal human observer performance while reducing inter-observer reproducibility eg (*11*), it might outperform existing human-based eg (*12*) or algorithm-based eg (*13*) performance (although this will have to be proven in prospective studies), or it might even uncover unknown phenomena eg (*14*). Whatever the scenario, the added-value ofthe 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 high-dimensionaldata. When AI is used to automate a task that humans can do by learning from manyexamples, the rules are deduced by the AI from the examples. The performance to be expected will depend on the representativity of the learning ensemble comparedto 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 anexample, we are currently at a loss when predicting why certain patients will respond to immunotherapy while 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, the misinterpretation of an AI decision-making process was published in highly respected journal (*15*), before a re-analysis of the data elegantly demonstrated the incorrect understanding of the initial results (*16*). This 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 regarding the biological mechanisms that are 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 / in vivo experimental design. Explainable AI is currently an extremely active area of research, with the ongoing development of numerous methods for approaching explainability[2], although fully satisfactory explanations may not always be feasible due to the high complexity anddimensionality of the data (*17*). The “Is it explainable?” question is thus certainly themost 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.

It is our conviction that the articles for which all four T.R.U.E. questions are convincingly addressed have a much higher likelihood of yielding significant advances in our field compared to 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 our devoted reviewers to keep thischecklist in mind when reviewing articles involving AI-algorithms. In addition, to further assist investigators in the developments of sound and reproducible AI-basedresearch, the Society of Nuclear Medicine and Molecular Imaging AI task force will soon release consensus recommendations underlying the specifics associated with Nuclear Medicine applications.

[2] https://christophm.github.io/interpretable-ml-book/

DISCLOSURE

No potential conflicts of interest relevant to this article exist.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their insightfulcomments and David Wallis for carefully proof-reading the manuscript.

REFERENCES

1. Sibille L, Seifert R, Avramovic N, et al. ^{18}F-FDG PET/CT uptake classification inlymphoma 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. *MedImage Anal*. 2020;63:101694.
5. Reuzé S, Orlhac F, Chargari C, et al. Prediction of cervical cancer recurrence usingtextural features extracted from 18F-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 inradiology. *Insights Imaging.* 2020;11:65.
8. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility inartificial intelligence. *Nature.* 2020;586:E14-E16.
9. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of imagemining: a systematic review on artificial intelligence and radiomics. *Eur J Nucl Med Mol Imaging.* 2019;46:2656-2672.

10. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance againsthealth-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digital Health.* 2019; 1: e271–97.

11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skincancer with deep neural networks. *Nature.* 2017; 542: 115-118.

12. 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. *NatMed*. 2019;25:954-961.

13. 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.

14. 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.

15. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*.2014;5:4006.

16. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signaturedevelopment: The need for safeguards. *Radiother Oncol*. 2019;130:2-9.

17. Bathaee Y. The artificial intelligence black box and the failure of intent andcausation. *Harvard J Law Tech*. 2018;31 :890-938.