EDITORIAL

Five principles for the intelligent use of AI in medical imaging

Errol Colak^{1,2*} , Robert Moreland^{3,4} and Marzyeh Ghassemi^{5,6}

© 2021 Springer-Verlag GmbH Germany, part of Springer Nature

Clinical scenario

You are a critical care physician at a medium-sized urban hospital. A local start-up is promoting software that claims to automatically interpret chest X-rays. Having read many articles about artificial intelligence, you are intrigued by the idea of using it. However, you have never had to decide if, when, and how to deploy these technologies in clinical practice.

Background

Artificial intelligence (AI) promises to transform health-care by improving clinical decision-making, reducing errors, and automating tasks [1]. Recent advances in machine learning (ML), a branch of AI, have demonstrated expert-level task performance including medical image analysis [2]. With an increasing number of medical imaging ML systems receiving regulatory approval, critical care physicians will have an additional set of tools at their disposal. However, the deployment of these systems into clinical environments can be fraught with pitfalls. This article introduces 5 key principles that should be remembered as the adoption of ML systems becomes more widespread.

Knowledge base: ensure you have enough understanding of machine learning to separate fact from fiction in vendor claims

If physicians are to be integral to the development, adoption, and oversight of ML models, it is paramount for them to develop basic competency in ML. Physicians should become familiar with dataset preparation, model architecture selection, and how statistical methods are applied

during model evaluation. Knowledge and skills in ML would empower physicians to understand the potential capabilities and limitations of ML, critically appraise and validate models, and address patient concerns. A deeper understanding of how ML models are created and evaluated may help to determine how model outputs can be best integrated into clinical decision-making. Readers are encouraged to explore the many online educational resources, many free of charge, that are available for self-directed and class-based learning (Supplemental Material).

Metrics: ensure that the evaluation metrics provided by the vendor are meaningful in clinical practice and not limited to accuracy

Model performance is often reported as an accuracy measure or confusion matrix. This is potentially problematic as models may demonstrate reduced accuracy when operating under real-world circumstances [3]. Furthermore, these metrics fail to consider prevalence, the impact of incorrect recommendations, and may in fact convey a false sense of safety. Determining a model's clinical benefit, robustness, direct and indirect costs, calibration, safety profile, interpretability, and user acceptance are much more meaningful forms of evaluation and are unfortunately often not available [4]. Rather than focusing on the area under the curve, a commonly used metric in the ML literature, the focus needs to be on rigorous clinical validation. The validation process for a novel ML model should be no less rigorous and thorough than that of a new drug or diagnostic test to ensure efficacy, reliability, acceptability, and safety [5].

*Correspondence: Errol.Colak@UnityHealth.to

¹ Department of Medical Imaging, Unity Health Toronto, University of Toronto, 30 Bond Street, Toronto, ON M5B 1W8, Canada
Full author information is available at the end of the article

Bias: ensure the data used for model development by the vendor reflects the case mix of your ICU and that the model has been validated using external datasets.

At their heart ML models are pattern recognition systems and can perform poorly outside the scope of the data used to train them [6]. They are highly dependent on the characteristics of this training data and vulnerable to any underlying selection biases. Specific patient populations and types of pathology may be under-represented during training but commonly encountered in your practice. Selection bias can reflect the patient population (e.g., age, gender, race), range of pathology (e.g., type and severity), and medical images (e.g., specific equipment, imaging protocol, patient positioning) used during training [7]. When evaluating an ML system, it is critical to understand the composition of the imaging studies used for training and whether it approximates those at your institution. Models trained on larger and more generalized datasets reduce the risks from bias but cannot entirely eliminate it.

Model fragility: ensure the vendor provides the parameters within which their model should operate safely.

The behavior of an ML system outside the scope of training data can be unpredictable and potentially unsafe [8]. For example, ML models may fail to detect obvious pathology if the imaging study presented is dissimilar to training data [9]. More disturbing is that ML models may fail to recognize that they are operating beyond their “comfort zone” and produce incorrect outputs with high confidence [10]. Also, ML systems are often trained to distill complex imaging data down to simple binary results: is the pathology present or not? This loss of granularity can result in cases close to the decision boundary being marked as normal even when clinically relevant findings may be present. When considering the implementation of an ML system, it is important to know the parameters within which the model should operate safely, if the model provides a measure of prediction confidence, and how the model will respond to the incomplete or poor-quality imaging studies that are often encountered in the real world.

Error-prone behavior: ensure the clinical team does not develop overdependence on the ML technology and loses sight of its fallibilities

Computer aids and automation are employed to improve efficiency and reduce error. However, they may actually lead to new and different types of errors [11]. Individuals may hold beliefs about the superiority of computer systems over human experts [12] and a consistently reliable system may even appear infallible. Blind trust in computer aids may impair an individual's

ability to recognize incorrect computer advice which may ultimately result in decision errors [11]. Furthermore, there is a risk of over-reliance on computer aids particularly in a clinical setting when many concurrent tasks compete for an individual's time and attention [13]. Critical care physicians must be aware that ML models are imperfect and can be used in biased ways; so, human judgment will be necessary in determining how much weight should be given to them in the clinical decision-making process.

Revisiting the clinical scenario

Having improved your understanding of AI systems, you realize the usefulness of the proposed system may be overstated. Similar to many current commercial systems targeting medical imaging, the start-up is marketing the product for widespread use but has not validated it on external data or in a full clinical setting [14]. Worse, its small training set consists mostly of PA radiographs of non-ICU patients which are notably different from your typical patients. Keenly aware that the value of the system is solely how it behaves in your institution, you insist on more thorough evaluation of the product with local imaging cases prior to further consideration.

Key points for the intelligent use of AI in medical imaging

To effectively evaluate a machine learning (ML) system, you should understand how they are created and the common methods used to assess them.

Proper evaluation of an ML system requires you to consider more than the tool's accuracy or ROC curve.

ML systems can be biased in a variety of ways and one that works well in one clinical environment may perform very poorly in another.

It is common for ML systems to behave unpredictably and potentially dangerously outside the scope of their training data.

Using an ML system can introduce new types of errors into the clinical decision process.

Electronic supplementary material

The online version of this article (<https://doi.org/10.1007/s00134-020-06316-8>) contains supplementary material, which is available to authorized users.

Author details

¹ Department of Medical Imaging, Unity Health Toronto, University of Toronto, 30 Bond Street, Toronto, ON M5B 1W8, Canada. ² Li Ka Shing Knowledge Institute, St. Michael's Hospital, 209 Victoria St, Toronto, ON M5C 3G7, Canada. ³ Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, 75 Francis St, Boston, MA 02115, USA. ⁴ MGH & BWH Centre for Clinical Data Science, 100 Cambridge St, Suite 1303, Boston, MA 02114, USA. ⁵ Department of Computer Science, University of Toronto, 214 College St, Toronto, ON M5T 3A1, Canada. ⁶ Vector Institute CIFAR AI Chair, 661 University Ave, Suite 710, Toronto, ON M5G 1M1, Canada.

Author contributions

All authors contributed to manuscript conceptualization, composition, and editing.

Funding

None.

Data availability**Not applicable.****Code availability****Not applicable.****Compliance with ethical standards****Conflicts of interest**

None.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 4 September 2020 Accepted: 29 October 2020

Published online: 15 January 2021

References

1. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 25(1):44–56
2. Majkowska A, Mittal S, Steiner DF et al (2020) Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* 294(2):421–431
3. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI conference on human factors in computing systems 2020;1–12
4. Nagendran M, Chen Y, Lovejoy CA et al (2020) Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 25:368
5. Park SH, Do KH, Choi JI et al (2018) Principles for evaluating the clinical implementation of novel digital healthcare devices. *J Korean Med Sci* 61(12):765–775
6. Brady A, Emanuele N (2020) Artificial intelligence in radiology—ethical considerations. *Diagnostics* 10(4):231
7. Oakden-Rayner L, Dunnmon J, et al. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM conference on health, inference, and learning. 2020;151–159
8. Cook T (2020) The Importance of Imaging Informatics and Informaticists in the Implementation of AI. *Acad Radiol* 27:113–116
9. Amodei D, Olah C, Steinhardt J, et al. (2016) Concrete problems in AI safety. arXiv preprint arXiv:1606.06565
10. Challen R, Denny J, Pitt M et al (2019) Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 28(3):231–237
11. Coiera E, Westbrook JI, Wyatt JC (2006) The safety and quality of decision support systems. *Yearb Med Inform* 15(01):20–25
12. Dijkstra JJ, Liebrand WB, Timminga E (1998) Persuasiveness of expert systems. *Behav Inf Technol* 17(3):155–163
13. Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. *Hum Factors* 52(3):381–410
14. Kim DW, Jang HY, Kim KW et al (2019) Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol* 20(3):405–410