International Journal of Computer Assisted Radiology and Surgery (2024) 19:609–611
https://doi.org/10.1007/s11548-024-03088-0

EDITORIAL

Regulation of AI algorithms for clinical decision support: a personal opinion

Kris Kandarpa[1]

Received: 14 February 2024 / Accepted: 27 February 2024 / Published online: 13 March 2024

Introduction

If information is the fabric our societies, then misinformation is the scissors. Misinformation in science and in medicine will destroy the trust borne of a long-existing assumed, but fragile, social contract. In abiding with this contract with their patients, physicians must responsibly apply their skills and tools, be it for diagnosis or treatment, by respecting the central maxim of medicine of first doing no harm. This should be the guiding principle for all stakeholders in this present context.

The stakeholders are many-patients, physicians, payors, regulators and front-and-center here, developers of AI clinical decision support [CDS] systems, which includes industry. Data scientists fully understand that errors propagate downstream. Nevertheless, everyone in the chain, which is only as strong as its weakest link, is responsible for avoiding unintentional adverse consequences created in the deployment of AI technologies. This opinion piece, admittedly non-comprehensive, will focus on the important role of regulators and their immediate upstream and downstream stakeholders—namely developers and physicians.

Developers and industry

Governments assign healthcare regulatory agencies with the authority to approve relevant marketable developments when the benefits far outweigh the risks to patients. In the process, regulators must assess the trustworthiness of CDS algorithms [CDSA]. Software developers must responsibly entrench credible quality controls along the entire developmental chain to assure trustworthiness. For training and validation of algorithms for an intended task, developers must collect AI-ready datasets that are ethically sourced [1], bias-free and representative of diverse patient populations [2], and that have been rigorously deidentified [if obtained from open data sources], curated and harmonized to the highest *acceptable* [or established] regulatory standards. Meticulous provenance of data [3], from whom, where [including imaging infrastructure and protocols], and when data were collected, along with documentation of population diversity [gender, racial, ethnic, socioeconomic, appropriate age distribution] within geographic areas should be non-negotiable. 'Time-stamped' datasets, being important for monitoring temporal drifts in the contained data [e.g., due to changes in disease prevalence, population migration, technology or standards of clinical care], would provide auditable trails to enable version control and replicability, while enhancing the generalizability of an algorithm.

Developers should transparently identify their data sources, the known limitations of their training and validation datasets, and provide brackets for the expected performance of the algorithm for the intended population and clinical task. CDSA developers can use available bias evaluation tools and quantifiable performance measurements to identify, assess and improve the limitations of algorithms [4].

A concrete method developers may use during deployment is to provide data source 'nutrition labels' [like content labels on packaged food] specifying the geographic diversity, representation [gender, age, racial, ethnic, socioeconomic, etc.] and collection periods of data used for training and validation [5]. Such ideas are reflected in the concept of *model cards*, which are already in limited use [6]. In addition, emulating 'drug sheets' that are provided in medicine packages, developers could declare intended 'indications and contraindications for use,' contextual efficacy [based in statistical data, e.g., ROC curves], and perhaps even identify potential risks from misapplication of the product.

Disclaimer: This invited editorial expresses the opinion of the author and does not reflect the positions or policies of the US NIH, US HHS, or US Government.

✉ Kris Kandarpa
kris.kandarpa@nih.gov

[1] Research Sciences and Strategy, National Institutes of Biomedical Imaging and Bioengineering, The US National Institutes of Health, Bethesda, MD 20892, USA

Regulators and governments

Although methods for improving generalizability of algorithms will not be discussed here, it is important to note that it will be limited if diversity and equitable representation of subpopulations within a region are neglected. Consequently, regulators may require developers to reveal the characteristics and limitations of the data source and declare optimal 'conditions' under which the submitted CDSA should be employed. This may include CDSA that are intentionally limited by design for use on specific subpopulations which have a high prevalence of the findings or disease. In consideration, regulatory agencies should provide guidelines based on *acceptable* standards that are tailored to the risk of the application, and that reflect the laws and policies of their nations.

Regulators, for their part, could develop their own pre-certification performance measures. Partnering with independent trusted institutions [e.g., such as Underwriters Labs in the USA], they could test algorithms submitted for certification on context-specific AI-ready datasets, sequestered from public use and accessible only to regulators [MIDRC.org]. Optimally, sequestered datasets would be created jointly by the regulatory agency and selected dataset hosts, and subject to agreed usage guidelines, to assure safeguards against potential 'gaming' by developers. Clearly, these data repository hosts would also have to rigorously abide by quality measures desired of developers. US NIH's MIDRC [a partnership between the RSNA, ACR and AAPM] is prospectively partnering with the US FDA on such a pilot project. Hypothetically, developer-submitted algorithms may be validated with AI-ready sequestered images [currently COVID-19 images only] as additional scrutiny for approval by the agency.

Regulators may also consider time-limited certifications, with periodic re-certification of CDS algorithms guided by performance metrics from real-world evidence, a practice already in use [7]. This post-market evaluation could be especially useful when there are temporal changes in disease prevalence, geographic shifts of populations or for expanded use to other populations as generalizability improves. As an example, the European Union has considered controlled limited dissemination of potentially high-risk algorithms. Such tempered measures can guide expanded approvals based upon real-world performance in assessing generalizability beyond the original claim.

In establishing regulatory laws, governing bodies must weigh the purpose, safety and efficacy of the use of health information technologies, while balancing cultural values pertinent to personal security and privacy. Consistent with this responsibility, the European Union's AI Act requires a 'classification system declaring level of risk' [8, 9]. The 'drug sheet' concept discussed earlier mirrors this in intention. In the USA, health information algorithms are regulated as medical devices—under the category of 'software as a medical device.' The US FDA regulates medical devices through three mechanisms—equivalence to existing devices [510[k]], de novo and [PMA] post-marketing approval [10]. Some criticize the 510[k] mechanism for approval based on 'self-declared similarity' because these approvals have had high post-market recalls. It has been reported that about a fifth of approved AI software marketing claims are not consistent with FDA clearance language [based on originally submitted claims] [7]. The US Algorithm Accountability Act [2019]—pertaining to high-risk automated decision systems [not just for health care]—requires an 'impact statement,' especially if personal information is involved [7]. These early exploratory attempts at 'regulation' of AI use in medical technology may become international and vastly more comprehensive and rigorous in the not too distant a future [8].

Other filters that regulatory agencies may use in their quest for trust in CDSA, especially those based upon deep learning are by emphasizing 'explainablity' and 'interpretability' [11, 12]. Developers can utilize the former to explain and improve algorithm function and the latter during implementation to help end-users to understand algorithm output. Early attempts to enhance 'interpretability' include *saliency* mapping [identification of data points that contributed to the diagnosis] and *ablation testing* whereby certain data points or features are deliberately removed to see how this impacts the performance of the algorithm. Other innovative methods are sure to arise.

Deep learning systems use convolutional neural networks [CNN], which emulate the biological brain. Ironically, humans comprehend neither CNN nor biological neural networks enough to understand how these systems reach their final decisions. Nevertheless, humans do subconsciously fall back on prior trusted truths—right or wrong—to come to new conclusions and can usually explain [defensible or not] their reasoning behind conclusions they draw. Computers certainly surpass humans in memory capacity and information processing speeds, so they should be able to chart and account of their 'reasoning' better than humans do. Demanding as it may seem, an algorithm should be required to convincingly self-declare its reasoning for arriving at a conclusion, potential pitfalls and limitations ['I have not been trained to answer this question'].

Physicians and healthcare systems

As already amply discussed, *clinical decision support* algorithms must be trusted by users. Gaining and maintaining user trust is an even greater hurdle for CDSA that employ 'black box' deep learning methods. This is not to imply

that physicians have always only worked with unambiguous medical information. Aspirin and 'medicinal' leeches were prescribed even while the *how* or *why* the 'treatments' worked was not understood. Subconscious confirmation bias [or a placebo effect] probably led practitioners to believe that the target ailment was indeed treated, if not cured by the intervention. However, through subsequent scientific enquiry, humans now better comprehend the mechanisms of action of the above practices [e.g., the anticoagulant molecule from leeches has been identified and synthesized into hirudin, an approved anticoagulant drug], giving us some comfort in prescribing them for an indicated use. Still, the question remains—should physicians accept deep learning systems whose opaque reasoning [black box] leads to an outwardly credible conclusion *just because* it confirms [potentially biased] expectations? Even if black box-generated answers are accepted for the time being, suspicion and skepticism should be the order of the day [13] until there is a convergence of the answers from 'black box' outputs and 'the truth to date' found through scientific enquiry.

Furthermore, testing and validation that are solely based on retrospective datasets while appropriate for preclinical development could cause liability in practice and should be unacceptable for clinical regulatory approvals. When the stakes are high with medical innovations whose benefit/risk ratio we do not fully know, clinicians and regulators must insist upon clinical trial testing and approvals must be based on convincing trial performance, like current medical regulatory practices for new medicines and devices [8, 14].

Conclusion: together, we swim, or we sink

All of us should adopt a buyer beware skepticism understanding that CDSA systems are not fool proof. Wisdom dictates that 'trust but verify' applies even to 'explained' conclusions. Unquestioning dependence on AI technologies, especially without comprehending their inner workings, will inevitably lead to disastrous consequences. For the foreseeable future, AI CDS technologies should be employed with *physician oversight* [14], especially for rendering a final diagnosis or deciding on an intervention. Today, it is appropriate to use AI-based workflow enhancing software, but it is too early in the gestation of deep learning technologies for them to override humans in establishing a diagnosis or treatment course, despite the introduction of large language model chatbots [15, 16]. But it is not too early for educating and training medical students, for that matter currently practicing healthcare workers, in the limitations and proper use of the AI systems of today and tomorrow [17, 18].

References

1. Herrington J, McCradden MD, Creel K et al (2023) Ethical considerations for artificial intelligence in medical imaging: data collection, development, and evaluation. J Nucl Med 64:1848–1854
2. Whitney HM, Baughan N, Myers K, et al. (2023) Longitudinal assessment of demographic representativeness in the medical imaging and data resource center open data commons. J Medical Imaging— in press. https://arvix.org/abs/2303.10501
3. Federov A, Longabaugh WJR, Pot D et al (2023) National cancer institute imaging data commons: towards transparency, reproducibility, and scalability in imaging artificial intelligence. Radiographics 43:1–14
4. Drukker K, Chen W, Gichoya J et al (2023) Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. J Med Imaging. https://doi.org/10.1117/1.JMI.10.6.061104
5. Mitchell M, Wu S, Zaldivar A, et al. (2019) Model cards for model reporting. Association of Computing Machinery. https://doi.org/10.1145/3287560.3287596
6. IBM AI Factsheet (2023). https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/actsheets-model-inventory.html?context=cpdaasf
7. Shah S and El-Sayed A. (2021) Medical Algorithms Need Better Regulation. Scientific American. www.springernature.com/us
8. Nelson, A. (2024) The right way to regulate AI. Focus on its possibilities, Not Its Perils. Foreign Affairs: https://www.foreignaffairs.com/united-states/right-way-regulate-aialondra-nelson
9. Thierer A and Chilson N. (2023) The problem with AI licensing & an FDA for algorithms. The Federalist Society. info@fedsoc.org.
10. FDA Clinical Decision Support Software (2022) Guidance for Industry and Food and Drug Administration Staff. Document issued on September 28
11. Fuhrman JD, Gorre N, Hu Q et al (2021) A review of explainable and interpretable AI applications in COVID-19 imaging. Med Phys. https://doi.org/10.1002/mp.15359
12. Cui S, Traverso A, Niraula D et al (2023) Interpretable Artificial Intelligence in Radiology & Radiation Oncology. Br J Radiol. https://doi.org/10.1259/bjr.20230142
13. Neeley T. (2023) Eight questions about using AI responsibly, answered. The big ideas series/ethics in the age of AI. (@tsedal) Harvard Business Review
14. Huag CJ, Drazen JM (2023) Artificial intelligence and machine learning in clinical medicine. N Engl J Med 388:1203–1208. https://doi.org/10.1056/NEJMra2302038
15. Lee P, Bubeck S, Petro J (2023) Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 388:1233–1239
16. 'A New Era' for Europe-EU AI Act Agreed: https://european-union.europa.eu/index_en
17. Moassefi M, Faghani S, Khosravi B et al (2023) Artificial intelligence in radiology: overview of application types, design, and challenges. Semin Roentgenol. https://doi.org/10.1053/j.ro.2023.01.005
18. AI in Health Professions Education: Proceedings of a Workshop (2023). National Academies. https://doi.org/10.17226/27174