Safeguards for the use of artificial intelligence and machine learning in global health

Amy K Paul[a] & Merrick Schaefer[b]

The potential benefits of artificial intelligence and machine learning are gaining attention in public health as in other fields. With applications spanning clinical decision support to management of supply-chain systems, artificial intelligence-enabled technologies are poised to improve clinical care and strengthen health systems.[1] Given the pace of progress in the application of such tools in advanced economies, we consider several challenges that low- and middle-income countries need to overcome to develop and deploy similar innovations. Health systems will have an important role in shaping the development of artificial intelligence-based tools and realizing their benefits. We argue that lower-resource countries will need to invest in improvements in data quality, improve equity in access to care, establish safeguards to minimize the harmful effects of bias and address supportive linkages within the health system. Without these developments, the benefits of the new technologies may fail to materialize, further exacerbating health disparities within and between high- and low-income countries.

Minimizing bias in training data sets

Potential applications of artificial intelligence in global health-service delivery include: improving health surveillance; enabling individuals to self-assess their health risks; equipping frontline health workers with tools for more accurate referrals, personalized interventions and diagnostic aids; and clinical decision support systems.[2] Developments in these artificial intelligence-enabled technologies, however, rely on large data sets to support the machine-learning algorithms from which artificial-learning tools excel at providing high-quality insights to very specific questions. However, these tools are only as good as the data used to train them.

The health sector is beginning to see the consequences of bias in the data used for machine training. For example, computer vision algorithms can classify images of skin lesions as malignant or benign to produce fast, accurate, non-invasive diagnoses.[3] However, image classifiers can perform differently on darker skins and therefore contribute to health disparities across populations.[4] Bias in dermatology data sets can stem from the input of smaller sample sizes from dark-skinned patients, or dark-skinned patients being identified at later stages of illness, both of which can lead to more frequent misdiagnosis. This type of bias can partly be addressed by ensuring training data are representative of the patients with whom such tools will be used. Yet this is not easy to fix. Many available data sets for training image classifiers, such as the International Skin Imaging Collaboration, have come from patients in Europe, North America and Australia, and do not reflect the diversity of patients or health conditions in lower-resource settings. The nuances of the context in which training data are collected become encoded in the results produced by artificial intelligence-based tools and therefore limit the ability of such technologies to work in different geographical, ethnic and economic contexts.

To produce valid results, many service-delivery applications based on machine learning will require the input of large amounts of patient-level data covering accurate diagnoses of the diseases and conditions that are common in lower-income settings. Applications aimed at prevention of illness and patients' retention in care need to reflect the local behaviour patterns of patients and to communicate the results in local languages and through culturally appropriate channels. Developing and scaling-up artificial intelligence-based innovations for use in low- and middle-income countries will thus require deliberate efforts to generate locally representative training data.

Improving system-level data

In the era of artificial intelligence, shortcomings in the quality, completeness and equity of health data generate particular risks.[5,6] Aggregated errors in data recording in lower-resource countries could culminate in misdirected interventions and resource allocation.[7] A machine-learning application intended, for example, to predict the likelihood that patients will default from treatment regimens requires accurate data on appointments. In health systems that cannot accurately share patient-level data across all facilities, it may be impossible to distinguish true defaulters from patients who seek care at different facilities. If the latter are treated as true defaulters, models produced by machine-learning algorithms may predict these patients to be high risk and misdirect resources to them. The results could potentially reinforce existing disparities in care within low-resource settings. Some novel approaches using machine learning have been applied to mitigate the underlying deficiencies of data from lower-income countries. Nevertheless, the long-term solution requires investments that improve the timeliness, accuracy, completeness, coverage and security of health data. Effective improvement of data collection requires strengthening health management information systems, while ensuring frontline workers have the training, support and capacity to do their work effectively.

[a] United States Agency for International Development, 1300 Pennsylvania Ave NW, Washington, DC, 20004, United States of America (USA).
[b] U.S. Global Development Lab, USAID, Washington, DC, USA.
Correspondence to Amy K Paul (email: apaul@usaid.gov).

(*Submitted: 15 May 2019 – Revised version received: 17 September 2019 – Accepted: 18 November 2019 – Published online: 27 January 2020*)

Strengthening health systems

Even when machine-learning tools produce valid results, their current capabilities, risk screening, diagnosis and identification of future threats, often provide only potentially actionable information. Information that informs only isolated interventions will usually not yield beneficial outcomes from the health-care response. In poorly functioning health systems, interventions of known efficacy may have little impact on outcomes because they depend on a cascade of intervention layers to work in combination.[8] In the case of a machine-learning diagnostic tool, the health-care worker may not know what to do with the diagnosis or may not have access to the products and services to treat it, while the patient may not have access to a health-care facility to obtain the necessary care. Digital health interventions are increasingly recognized as discrete strategies for overcoming bottlenecks within a health system rather than solutions that directly cause outcomes. Artificial intelligence and machine-learning tools are no different.[9]

Ensuring trust

Failure to address these system-level bottlenecks may not only fail to achieve the desired outcomes from the use of artificial intelligence-based tools in low- and middle-income countries, but also pose threats to the trust that underpins a functioning health system.[10] Machine-learning tools cannot fulfil their purpose if the data input during development are not representative of the context where the tools are applied, or if poor-quality training data lead to misleading conclusions, or if the recommended action is not possible due to health-system constraints. Whatever the reason, the patient perceives only a failure to receive the needed care. In this way, artificial intelligence tools are just as fallible, and hence detrimental to trust, as other digital tools that are widely adopted before we have understood whether, and in which contexts, they work.

The inherent complexity of artificial intelligence-based tools presents additional risks of loss of trust. For tools such as the melanoma classifier, the complexity of many image classification algorithms (based on artificial neural networks) makes it nearly impossible for humans to understand how a classification was made and ensure they are not relying on spurious associations.[11] If clinicians are unable to follow the diagnostic pathway of machine-learning models, and hence are unable to review the diagnoses generated, there is an increased risk of failing to identify and correct biased outcomes when they do occur.

Concerns have also been raised about how humans react to failures in artificial intelligence systems.[12] In global health care, the failures of artificial intelligence-based tools will likely stem from the technological complexity, in addition to issues of demand, access, coverage, quality and affordability within health systems. Yet when such tools fail, there is a risk that frontline workers are viewed as responsible. One author has described the concept of moral crumple zones in which the people acting on the results of complex technological systems have to absorb the blame for failures, just as the crumple zones of automobiles absorb the brunt of the force in collisions.[12] People tend to retain trust in the systems that produce complex technological tools at the expense of trust in the people who use them. When applying new artificial intelligence-based tools in low-resource health systems, technological failure could undermine the trust in the people and institutions that are needed to fully realize the potential of artificial intelligence.

Preparing health systems

Given the multiple ways in which tools based on machine learning may fail, we need a strategic approach to investments in artificial intelligence for global health services. Investments are needed that strengthen health systems and support the development of relevant, accurate solutions that work for the diversity of populations who need them. Now is the time to prioritize health-system investments that will: (i) improve the quality (completeness, accuracy and representativeness) and use of locally generated data through investments in health management information systems; (ii) increase the representation of poor and marginalized groups in the training data used to develop machine-learning based tools by improving equity in patient access to health care; (iii) establish safeguards against bias, such as standards for ensuring representativeness and transparency of training data sets and processes for interrogating how automated clinical-decision support tools work; and (iv) invest in machine-learning tools only in contexts where the health system is strong enough to support converting the results into action.

Such investments will not only mitigate the risk of bias and poor performance in artificial intelligence-based tools, but also build and preserve trust in the health systems needed to realize their benefits. If these tools are to succeed at scale, it will not be despite the underlying health systems, but because of them. ■

Competing interests: None declared.

References

1. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. AI and global health: how can artificial intelligence contribute to health in resource-poor settings? BMJ Glob Health. 2018 08 29;3(4):e000798. doi: http://dx.doi.org/10.1136/bmjgh-2018-000798 PMID: 30233828
2. Artificial intelligence in global health: defining a collective path forward. Washington, DC: United States Agency for International Development; 2019. Available from: https://www.usaid.gov/cii/ai-in-global-health [cited 2019 Sep 16].
3. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al.; Reader study level-I and level-II Groups. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018 08 1;29(8):1836–42. doi: http://dx.doi.org/10.1093/annonc/mdy166 PMID: 29846502
4. Adamson AS, Smith A. Machine learning and healthcare disparities in dermatology. JAMA Dermatol. 2018 11 1;154(11):1247–8. doi: http://dx.doi.org/10.1001/jamadermatol.2018.2348 PMID: 30073260
5. Wyber R, Vaillancourt S, Perry W, Mannava P, Folaranmi T, Celi LA. Big data in global health: improving health in low- and middle-income countries. Bull World Health Organ. 2015 Mar 1;93(3):203–8. doi: http://dx.doi.org/10.2471/BLT.14.139022 PMID: 25767300

6. Sundeep S, Sundararaman T, Braa J. Public health informatics: designing for change – a developing country perspective. New York: Oxford University Press; 2017.
7. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018 11 1;178(11):1544–7. doi: http://dx.doi.org/10.1001/jamainternmed.2018.3763 PMID: 30128552
8. Tanahashi T. Health service coverage and its evaluation. Bull World Health Organ. 1978;56(2):295–303. PMID: 96953
9. Mehl G, Labrique A. Prioritizing integrated mHealth strategies for universal health coverage. Science. 2014 Sep 12;345(6202):1284–7. doi: http://dx.doi.org/10.1126/science.1258926 PMID: 25214614
10. Gilson L. Trust and the development of health care as a social institution. Soc Sci Med. 2003 Apr;56(7):1453–68. doi: http://dx.doi.org/10.1016/S0277-9536(02)00142-9 PMID: 12614697
11. Bissoto A, Fornaciali M, Valle E, Avila S. (De)constructing bias in skin lesion datasets [online]. Paper presented at the 2019 ISIC Skin Image Analysis Workshop at the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, United States of America, 16–20 June 2019. arXiv. 2019 Apr 18:1904.08818v1. Available from: https://arxiv.org/pdf/1904.08818.pdf [cited 2019 Sep 15].
12. Elish M. Moral crumple zones: cautionary tales in human–robot interaction. engaging science. Technol Soc. 2019;5:40–60.