WHO and ITU establish benchmarking process for artificial intelligence in health

AI for Health
An ITU Focus Group
In partnership with WHO

Published Online
March 29, 2019
http://dx.doi.org/10.1016/S0140-6736(19)30762-7

Growing populations, demographic changes, and a shortage of health practitioners have placed pressures on the health-care sector. In parallel, increasing amounts of digital health data and information have become available. Artificial intelligence (AI) models that learn from these large datasets are in development and have the potential to assist with pattern recognition and classification problems in medicine—for example, early detection, diagnosis, and medical decision making.[1,2] These advances promise to improve health care for patients and provide much-needed support for medical practitioners.

Over the past decade, considerable resources have been allocated to exploring the use of AI for health. Although there is immense potential, many issues such as regulation, potential for bias, and adequate evaluation of efficacy must first be addressed for safe and ethical implementation of AI in health care.[3]

Modern AI algorithms are complex, and their performance depends on the quality of the training data and learning mechanism. If AI algorithms are poorly designed or the training data are biased or incomplete, errors can occur. There is no agreed framework for assessing or reporting the results of health AI models before deciding whether they are sufficiently robust for application in a population, as there is for new drugs or surgical interventions. The absence of confidence or quality control is a major barrier to the uptake of AI in health care. Creating a rigorous, standardised evaluation framework that leverages the advantages and addresses the limitations of AI models in health is crucial for realising the potential of this technology and limiting risks.

Two UN agencies, WHO and the International Telecommunication Union (ITU), established a Focus Group on Artificial Intelligence for Health (FG-AI4H) in July, 2018. FG-AI4H is developing a benchmarking process for health AI models that can act as an international, independent, standard evaluation framework.

To establish this evaluation and benchmarking process, FG-AI4H is calling for participation from medical, public health, AI, data analytics, and policy experts. Topic groups are being formed by communities of stakeholders allowing FG-AI4H to develop its processes for AI evaluation and benchmarking specific for each health topic. Each topic use case will be reviewed for its relevance and should impact a large and diverse part of the global population or solve a health problem that is difficult or expensive. The AI models are expected to offer improvements over current practices in quality or efficiency that would be expected to lead to better health outcomes or cost-effectiveness. Once formed, topic groups will provide a forum for open collaboration among stakeholders who agree on a pragmatic, best-practice approach for benchmarking each use case, including defining the application scenario and desired output of AI models in that use case, identifying adequate sources of training and testing data, and facilitating the preparation of multisource heterogeneous data. All data for training and testing are expected to be of high quality, ethically generated, and accompanied by detailed information about their format and properties. Thus far, FG-AI4H has developed 11 topic groups in areas such as cardiovascular disease risk prediction, ophthalmology (retinal imaging diagnostics), and AI-based symptom checkers, but this approach is expected to be expanded to other tasks.

For **FG-AI4H** see https://www.itu.int/go/fgai4h

The benchmarking process will be done on secure, confidential test data. Ideally, test data will originate from various sources to determine whether the use of an AI model can be generalised across different populations, measurement devices, and health-care settings. The benchmarking process for each use case within a topic needs to be defined. For many use cases it would, at least initially, be meaningful to compare model performance against human performance, or human performance with AI assistance in the same task, whereas for other tasks, comparative performance of algorithms would be more meaningful. Once these requirements are met, AI models can be submitted via an online platform to be evaluated with the test data. Established in this way, the benchmarking process will not only provide a reliable, robust, and independent evaluation system that can demonstrate the quality of AI models, but will also provide an independent test dataset for model validation consistent with best-practice recommendations for reporting multivariable prediction models in health.[4]

FG-AI4H has held three workshops and meetings and its ambitious task has received a positive reception from companies, academics, policy makers, software developers, and device manufacturers. The group will meet again in Shanghai, China, on April 2–5, 2019, with representatives from ITU and WHO in attendance. Further meetings this year are planned in Geneva, Tanzania, and India. We invite the academia, technology, and regulatory communities to contribute to FG-AI4H by sharing topics, data, expertise, use cases, and algorithms. The creation of an open and transparent process for evaluation of health AI models is key to realising the potential of AI to improve human health worldwide.

Thomas Wiegand, Ramesh Krishnamurthy, Monique Kuglitsch, Naomi Lee, Sameer Pujari, Marcel Salathé, Markus Wenzel, Shan Xu
Fraunhofer Heinrich Hertz Institute and Technische Universität Berlin, Berlin 10587, Germany (TW, MK, MW); World Health Organization, Geneva, Switzerland (RK, SP); *The Lancet*, London, UK (NL); École Polytechnique Fédérale de Lausanne, Geneva, Switzerland (MS); and China Academy of Information and Communications Technology, Beijing, China (SX)
thomas.wiegand@hhi.fraunhofer.de

We are all members of the WHO–ITU Focus Group on Artificial Intelligence for Health (FG-AI4H) and declare no other competing interests.

1 Topol E. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25:** 44–56.

2 Wahl B, Cossy-Gantner A, Germann S, Schwalbe N. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health* 2018; **3:** e000798.

3 Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. *Science* 2019; **363:** 810–12.

4 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015; **162:** 55–63.