# Walking the tightrope of artificial intelligence guidelines in clinical practice

Over the past few months, there has been a wave of digital health guidelines and whitepapers issued by regulators, institutes, and organisations worldwide. In the field of artificial intelligence (AI), EU guidelines, published in April, promote the development of trustworthy AI across all disciplines, while a US Food and Drug Administration (FDA) whitepaper proposes a regulatory framework for constantly developing software in health care. Guidelines from the National Institution of Health and Care Excellence (NICE) tackle the level of evidence required for a new digital health intervention, and NHSX and Public Health England have both reported their intention to produce their own AI guidelines.

AI approaches in medical practice needs to be lawful, ethical, and robust. According to the EU guidelines for trustworthy AI, there are seven key requirements for ethical AI: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination, and fairness; societal and environmental wellbeing; and accountability. They include tiered, risk-based guidance for tool validation for prevention of harm, recommendations to make the model explainable as well as fair and unbiased, and ensure that human autonomy is maintained. The guidelines highlight that AI approaches should augment the actions of humans through transparent decision pathways rather than black box decision making.

Assuming that a model has been designed and created ethically, what is the minimum level of evidence needed to use the AI in the clinic? It will, and should, vary depending on the function for which the software has been designed, which is recognised in NICE's guidelines for digital health interventions. An AI that recommends a dietary programme in people at risk of developing high blood pressure should require a different level of evidence to a program that recommends treatment options for patients in intensive care. Rather than AI-specific guidelines, NICE's recommendations are across all digital health interventions; as a result, AI-specific guidelines are not yet detailed enough to define which level of evidence is needed in each classification. In other disciplines, AI algorithms have been tested and validated primarily from the same source dataset. Through a random separation into training and test datasets, and cross validation to improve its reliability, the AI tool could be considered sufficiently reliable to be used in real-world settings with the potential to learn and improve with access to more data over time. However, for tools suggesting treatments and diagnosis algorithms, the models must be more robust to ensure patient safety. AI algorithms need to be trained on an independent and diverse validation dataset to confirm the effectiveness and generalisability of the algorithm. Validation of algorithms hinge on the availability of external data from public datasets.

By definition, an AI model is constantly evolving. The final consideration, therefore, is how to regulate these inevitable changes to AI models after approval for use in the clinic has been granted. This is addressed by the FDA whitepaper on modifications to software using machine learning models. Though not yet formal guidelines, the framework that has been issued for discussion is thoughtful and identifies three main areas under which the AI can change: performance, input, and intended use of the software. The last of these changes could be grounds for restarting the approval process, whereas other modifications need only be documented and subject to review periodically.

Guidelines securing the minimum level of clinical evidence required for different tiers of AI studies are necessary to eliminate variation in the quality of published studies and in the AI tools themselves. The existing guidelines discussed here and those still in development need to be updated frequently to ensure they remain relevant to advancing technologies. We are awaiting the publication of TRIPOD-ML reporting guidelines, which are being developed specifically for some issues that arise in AI studies published in journals such as ours. We will continue to require independent validation for all AI studies that screen, treat, or diagnose disease. Data should be diverse so as minimise bias and be of high quality to ensure veracity of the findings. We believe that the evidence threshold will evolve as technology advances to strive for increasingly accurate AI models in health care.

■ *The Lancet Digital Health*



Mohamed E. Abazeed

For the **NHSX policy guidance announcement** see https://www.publictechnology.net/articles/news/nhsx-create-policy-guide-use-ai-healthcare

For the **Public Health England guidance for the use of AI in screening** see https://phescreening.blog.gov.uk/2019/03/14/new-guidance-for-ai-in-screening

For the **EU ethics guidelines for trustworthy AI** see https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

For the **NHS code of conduct** see https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology

For **NICE's guidelines for digital health interventions** see https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf

For the **FDA whitepaper for modifications to software using machine learning models** see https://www.regulations.gov/document?D=FDA-2019-N-1185-0001

For **TRIPOD-ML reporting guideline statement** see **Comment** *Lancet* 2019; **393:** 1577–79