The Need for Targeted Labeling of Machine Learning–Based Software as a Medical Device

Benjamin A. Goldstein, PhD; Maciej A. Mazurowski, PhD; Cheng Li, PhD

Machine learning (ML)-based clinical decision support (CDS) tools are increasingly part of the health care landscape. These tools have the potential to automatically identify patterns and assign health risk in ways that human medical practitioners are not capable of doing. If implemented correctly, this has the potential to provide a more optimized and less expensive health care. However, there is also a downside where end users—who do not necessarily understand how the underlying ML algorithms operate—need to trust the accuracy of such tools. As the recent paper by Hsu et al\(^1\) reported, sometimes this trust may be unwarranted.

ML models for medical imaging has been one of the success stories of ML algorithms.\(^2\) Medical images consist of complex patterns. These patterns can be difficult for humans to decipher; however, ML models can be trained to disentangle them. Imaging AI models have been successfully applied in a variety of imaging modalities. Mammography has been an early application of ML with a CDS system approved by the Food and Drug Administration (FDA) in 1998. Since then, the effectiveness of the early CDS systems were put into question.\(^3\) This was followed by substantial technical advances in the field with the introduction of deep learning and its tremendous success for image analysis.

A common way to evaluate ML algorithms is to split the available population into a training set and test set, develop the model using the training set, and evaluate it using the test set. This type of evaluation, however, leaves out the question on how the model will behave for a population with different characteristics and images acquired using different scanners. Hsu et al,\(^1\) following an increasing number of studies of this type, followed a different type of evaluation. They took an ensemble of ML algorithms for mammography initially developed using data from the DREAM Mammography Challenge and evaluated it on an independent large data set with different demographic and clinical characteristics. The authors found the models performed worse in some subsets of women (eg, women with dense breast tissue and those with a history of breast cancer). In particular, they found that the models performed worse than radiologist-read images.

This raises important questions for how we consider using such tools in a clinical environment. The FDA has created draft guidance for software as a medical device. A tool used to diagnose breast cancer based on a mammogram would be considered a medical device and would require regulatory approval.\(^4\) Part of the regulatory framework for all medical devices is labeling the intended population of use. As the present study suggests, this is critically important. When the training and testing populations of an ML CDS differ from the application population, performance is no longer guaranteed. Within the regulatory process, as part of the final clinical validation step, the output data should be analyzed to ensure they meet the intended purpose of software as a medical device in the target population and in the context of clinical care. However, there is no specification of the extent to which a tool needs to match its intended population. As the draft guidance frames, there are areas where software is different from typical medical devices, necessitating different regulatory requirements (eg, clinical evaluation). As such, there is an opportunity to push this labeling requirement further. One unique aspect of ML tools is the opportunity to test the performance of the tool in a large range of subpopulations. Therefore, instead of labeling for the tool’s use in broad categories such as adult vs pediatric populations, developers can retrospectively test their software in more targeted populations such as women with dense breast tissue and label their tools.
appropriately. As seen in the paper by Hsu and colleagues, such targeted assessments are both feasible and vital.

In response to this, institutions, such as Duke University, are establishing governance frameworks for deploying CDS tools. A key aspect of our process is the local testing of CDS tools and assessing their performance across and within diverse patient populations before deployment. We view this as an important step to ensuring the high quality and equitable performance of any CDS. However, this raises additional challenges, as most health care environments do not have the resources to test out all (or even any) models locally. As such, it is important for local adopters to be cognizant of on whom an ML tool was developed, and how that population may meaningfully differ from one's local environment. Greater specification with labeling requirements for software as a medical device would aid this process.

We acknowledge that performing such targeted testing is not always simple. One of the biggest barriers is the access to diverse data sets. Currently, the field is dominated by benchmark data sets that do not always reflect the demographic and clinical diversity of intended use populations. The new requirements to make data supported by NIH studies publicly available would facilitate greater access to diverse data. As we have written previously, such data can be stored in cloud environments allowing both developers and end users to assess how their tools perform in diverse patient populations.

Ultimately, the integration of ML tools within clinical practice is upon us. The FDA has created a framework for adapting current medical device regulation for ML-based software. As the work of Hsu et al. show, there is a pressing need for more evaluation of these tools beyond standardized benchmark data sets, and into clinical settings that reflect real patient diversity. Moreover, there is a need to generate targeted labels to allow end users to understand for whom a tool is most appropriate. Applying a tool beyond these groups should require additional evaluation. Health systems can aid in this process by creating localized evaluation frameworks, to the extent resources and expertise allow. Ultimately, CDS tools have transformative potential for health care, but only if the tools are trustworthy and equitable. Better labels of intended use have the potential to aid in this regard.

ARTICLE INFORMATION
Published: November 21, 2022. doi:10.1001/jamanetworkopen.2022.42351
Open Access: This is an open access article distributed under the terms of the CC-BY License. © 2022 Goldstein BA et al. JAMA Network Open.

Corresponding Author: Benjamin A. Goldstein, PhD, Department of Biostatics and Bioinformatics, Duke University School of Medicine, 2424 Erwin Rd, Ste 9023, Durham, NC 27705 (ben.goldstein@duke.edu).

Author Affiliations: Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina (Goldstein, Mazurowski); Duke AI Health, Duke University School of Medicine, Durham, North Carolina (Goldstein, Mazurowski); Department of Radiology, Duke University School of Medicine, Durham, North Carolina (Mazurowski); Independent Regulatory Consultant, Durham, North Carolina (Li).

Conflict of Interest Disclosures: Dr Mazurowski reported receiving grants from Astra Zeneca, Aravive, Akesobio, GSK, Merck, Seattle Genetics research funds paid to institution outside the submitted work; and Dr Mazurowski reported a patent for systems and methods for extracting prognostic image features issued. No other disclosures were reported.

REFERENCES


