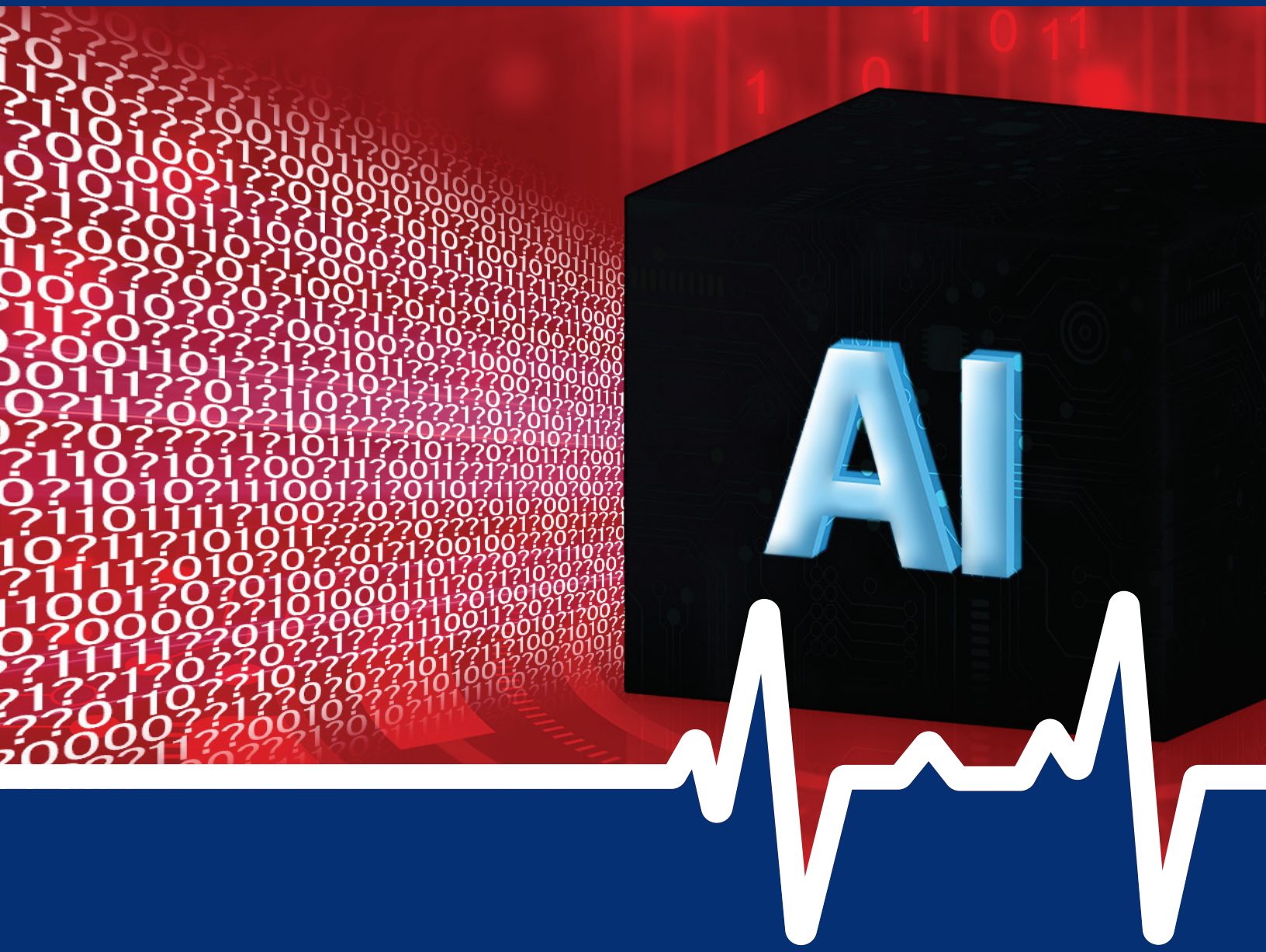


Generalization Uncertainty in AI-Enabled Medical Devices

A Safer Way Forward

Kev Coleman

PARAGON HEALTH INSTITUTE

MAY 2026

The author is especially grateful to Michael Pencina, Lee Fleisher, and other scholars for their thoughtful feedback regarding this paper. The final publication is a far superior work because of the many insights they shared and issues they raised. The author would also like to thank the Paragon team for their exceptional comments and work in review of the paper.

ABOUT THE AUTHOR

Kev Coleman oversees the Health Care AI Initiative at Paragon Health Institute. He is recognized as one of the leading experts in healthcare AI policy. His work has been featured by *Politico* and Fierce Healthcare and has shaped industry discussions on the safety and regulation of healthcare AI. He speaks frequently on this topic, having participated at events such as The Asian Leadership Summit in Seoul, South Korea, as well as Politico's 2025 Health Care Summit, where he presented his work during a panel discussion alongside Rep. Ami Bera (D-CA). His notable papers on AI include *Targeted Postmarket Surveillance: The Way Toward Responsible AI Innovation in Health Care*, *Healthcare AI Regulation: Guidelines for Maintaining Public Safety and Innovation*, and *Lowering Health Care Costs Through AI: The Possibilities and Barriers*.

Mr. Coleman's career spans work as a successful technology startup executive to healthcare research, and the results of the latter have been cited in top newspapers and media across the country and referenced in congressional health reform discussions (both Democratic and Republican).

EXECUTIVE SUMMARY

Artificial intelligence (AI) medical devices often perform well in testing but less reliably when used on real patients whose data differs from the data used to develop the model. Device unreliability presents patient safety risks and can erode clinician trust as well as slow the technology's adoption. From a policy perspective, unreliability may prompt regulatory responses that inadvertently constrain the benefits AI can bring to health care without solving the underlying problem. Thus, "generalization" is a crucial matter for AI-enabled medical devices.

Generalization is the technical term for an AI medical device's successful processing of real-world data (RWD) encountered in a medical facility. The result of this processing is called the device's output, and outputs differ according to device type. An output may be a disease prediction, a categorization (such as a diagnosis based on a medical image), a recommendation of a specific medical intervention, or some other clinical determination.

Generalization, however, is not a foregone conclusion for AI devices. Unlike traditional software systems that employ deterministic rules (e.g., "If x is present then do y, otherwise do z"), AI often uses complex models to predict the most likely output. The effectiveness of a model is closely related to the training data that configured the model's parameters and, in turn, determines the device's real-world performance. The central policy challenge is how to respond best to this uncertainty without impairing the field's ongoing progress or mandating an ineffective remedy.

Generalization uncertainty is doubt regarding a device's ability to produce accurate outputs. Research has found that many devices approved by the Food and Drug Administration lack robust clinical performance evaluations. Moreover, a 2025 study found that AI medical devices lacking validation were more likely to be recalled.[1] These validation gaps, compounded by the "black box" nature of AI, heighten concerns about reliability.

1 Branden Lee et al., "Early Recalls and Clinical Validation Gaps in Artificial Intelligence–Enabled Medical Devices," *JAMA Health Forum* 6, no. 8 (August 22, 2025), https://jamanetwork.com/journals/jama-health-forum/fullarticle/2837802.

As mentioned at the outset, generalization problems pose a genuine threat to patient safety. Alongside this risk of patient harm, generalization problems may also erode confidence in these devices' reliability and provoke policy reactions that depress AI device adoption. Inasmuch as AI technology is contributing to life-saving innovations — such as predicting when a patient is at high risk for breast cancer in the next 24 months[2] — generalization uncertainty could prove detrimental to American health care beyond individual device failures.

A worry related to generalization uncertainty is the possibility of algorithmic bias and, potentially, poor device performance for populations that are not well represented within device training data. However, the remedy for generalization uncertainty transcends adequate demographic representation, because AI medical devices are parameterized by specific features contained within their training data. If, for example, the training data is demographically diverse but the characteristics within individual data samples are very similar to one another, then outlier patients (who significantly deviate from these characteristics) may be at higher risk for device non-generalization despite belonging to the demographic segments adequately represented within the training data.

Among the proposed remedies for generalization uncertainty are third-party device and algorithm certification, training data assessment, and physician evaluation of training data suitability. Among the limitations observed among these options are the risk of high consultative costs, a conflict with future adaptive AI designs, and a failure to personalize the generalization question for an individual patient. High consultative costs are a particular concern, because they could encourage a divide between well-financed health systems that can afford such consultation and rural health systems that cannot.

Instead of mandating any of these options, this paper proposes the development of a voluntary alternative, Digital Similarity Analysis (DSA). DSA will evaluate the similarity and dissimilarity of an individual patient's medical image to the training and testing data used for the device's development. The purpose of DSA is to determine if a patient's medical image is an outlier compared to a device's training and testing data. This determination would be made *prior to the use of that device with the patient image*. The physician, when alerted by DSA that a patient's image is significantly dissimilar to the AI device's training and testing data, can decide to:

2 Disease prediction with AI is not as simple as "You will get disease x" or "You will not get disease x." Instead, AI may make a prediction across a time span, such as five years, and express the likelihood of disease occurrence in each of the years along with a confidence level associated with each prediction's accuracy. Related to AI's disease prediction capabilities is the potential of the technology to detect signs of disease before radiologists can visually perceive those signs. The advantage for both prediction and very early detection is medically intervening while the disease is more responsive to treatment. Additionally, such interventions may be less costly and maintain a higher quality of life for the patient during treatment. See May Kekatos, "How Artificial Intelligence Is Being Used to Detect, Treat Cancer — and the Potential Risks for Patients," *ABC News*, July 21, 2023, https://abcnews.com/Health/ai-detect-treat-cancer-potential-risks-patients/story.

- forgo device use because of the perceived risk for generalization problems,
- require supplemental validation of the medical device's output for the patient, or
- use the device but treat any clinical determination produced by the device with lower confidence.

Although the DSA proposal would not eliminate generalization uncertainty, it provides a valuable direction for AI medical device safety and avoids alternatives that inadequately address the problem. Further, DSA expands the discussion of algorithmic bias beyond broad demographic categories to the specific characteristics of each patient. By shifting evaluation from population groups to individuals, the DSA approach may enhance safety across demographic segments. Finally, an additional benefit of DSA is its integration of image characteristics that arise from differences in radiology equipment and technician technique, a subject too often ignored in the discussion of generalization uncertainty.

INTRODUCTION: AI GENERALIZATION AND BIAS CONCERNS

Artificial intelligence (AI) in health care has expanded at a rapid pace, advancing through a diverse range of applications that include transcribing physician-patient interactions, automating billing code assignments for patient care, and designing new drugs. With respect to AI in medical devices,[3] adoption has faced challenges related to their generalization abilities.

Generalization is the technical term for an AI-enabled medical device's successful processing of real-world data (RWD) encountered in a medical facility. The result produced from this processing is called the device's output, and outputs differ according to device type. An output may be a disease prediction, a categorization (such as a diagnosis), a recommendation of a specific medical intervention, or some other clinical determination. Generalization, however, is not a foregone conclusion for AI devices. Unlike traditional software systems that use rules to process RWD (e.g., "If *x* is present then do *y*, otherwise do *z*"), AI typically uses complex models[4] to predict the most likely output. The effectiveness of any model is closely related to the training data that configured the model's parameters[5] and, in turn, determined the device's real-world performance. If these parameters are configured too closely to the model's training data, then the device will likely have generalization deficiencies and the model inside the device will be described as overfit. A generalization deficiency is a technical way of stating that the output produced by an AI-enabled medical device is unreliable, either inaccurate or inadequate. Underfitting, in contrast to overfitting, occurs when the training data is insufficiently expansive[6] to address the wide variety of data scenarios expected in real-world use of the device.[7] Both overfitting and underfitting can produce generalization deficiencies — that is, the device's output is either inaccurate or inadequate.

Generalization uncertainty is doubt regarding an AI device's ability to produce accurate outputs when using RWD. Generalization problems, regardless of overfitting or underfitting, can pose a genuine threat to patient safety. Alongside this risk of patient harm, generalization

3 The Food and Drug Administration (FDA) — following the Food, Drug, and Cosmetic Act — classifies a product as a medical device if, among other features, it is used "in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease." FDA, "How to Determine If Your Product Is a Medical Device," September 29, 2022, https://www.fda.gov/medical-devices/classify-your-medical-device/how-determine-if-your-product-medical-device.

4 An AI model is the totality of an AI system's algorithms and their respective parameterizations. The model is what performs the principal functionality (e.g., diagnosis, prediction, etc.) within an AI-enabled medical device.

5 A parameter is a variable (i.e., a value such as a number or a string of letters) used to control software functions and performance.

6 *Expansive*, in this context, not only includes the diversity of patient data possibilities but also the range of effects that imaging equipment (e.g., MRI machines, x-rays, ultrasounds, etc.) can have on the form of that data. For example, newer imaging equipment may exhibit greater color and contrast sensitivity compared to older models and, when compared with the output of older models, produce image differences that are not due to patient factors. Technician technique in using a device (e.g. a diagnostic ultrasound machine) is another nonpatient factor that can produce image differences.

7 Underfitting can also occur if an AI model's algorithms are too simplistic.

problems may also erode confidence in these devices' reliability and provoke policy reactions that depress AI device adoption.[8] Because AI technology is contributing to life-saving innovations, such as predicting when a patient is at high risk for breast cancer in the next 24 months,[9] generalization uncertainty (and the AI adoption resistance it engenders) could prove detrimental to American health care beyond individual device failures.

Generalization uncertainty is a growing concern in clinical AI,[10] particularly given current deficits in device validations. A study of 903 AI-enabled medical devices approved by the Food and Drug Administration (FDA) found that "at the time of regulatory approval, clinical performance studies were reported for approximately half of the analyzed devices, while one-quarter explicitly stated that no such studies had been conducted."[11] A separate study of a similar number of AI medical devices found that 6.3 percent of the devices were subject to recall actions after FDA approval and that devices lacking either prospective or retrospective validation had more recalls per device.[12] The combination of validation gaps and recall data, along with the black box nature of AI devices, feeds generalization uncertainty. This state of affairs has given rise to the observation that "despite the promising potential and broad applications, transparent information regarding key characteristics and outcomes related to the clinical evaluation of these devices at the time of regulatory approval is not well established."[13]

Generalization uncertainty risks include algorithmic bias[14] and, more specifically, poor device performance for minority populations who are not well represented within device training data.[15] This worry is part of a larger concern about bias in medicine. Beyond the discussion of

8 Regarding the relationship between AI-enabled medical device reliability and clinical adoption, see Moustafa Abdelwanis et al., "Artificial Intelligence Adoption Challenges from Healthcare Providers' Perspectives: A Comprehensive Review and Future Directions," *Safety Science* 193 (January 2026), https://www.sciencedirect.com/science/article/pii/S092575352500253X#b99; and Masooma Hassan et al., "Barriers to and Facilitators of Artificial Intelligence Adoption in Health Care: Scoping Review," *JMIR Human Factors* 11 (August 29, 2024), https://pmc.ncbi.nlm.nih.gov/articles/PMC11393514/.

9 See footnote 1.

10 Lea Geotz et al., "Generalization — a Key Challenge for Responsible AI in Patient-Facing Clinical Applications," *NPJ Digital Medicine* 7, no. 126 (May 21, 2024), https://www.nature.com/articles/s41746-024-01127-3.

11 Daniel Windecker, "Generalizability of FDA-Approved AI-Enabled Medical Devices for Clinical Use," *JAMA Network Open* 8, no. 4 (April 30, 2025), https://pmc.ncbi.nlm.nih.gov/articles/PMC12044510/.

12 "Among 950 AIMDs, 60 (6.3%) were associated with 182 recall events.... Diagnostic or measurement errors accounted for 109 recalls encompassing 935063 units, followed by functionality delay (44 recalls, 755647 units), physical hazards (14 recalls, 8192 units), and biochemical hazards (13 recalls, 76257 units)." Lee et al., "Early Recalls and Clinical Validation Gaps."

13 Windecker, "Generalizability."

14 Representative of this concern is a 2025 remark published in *The Lancet*: "Artificial intelligence (AI) technologies come with many promises and risks. A notable and well-established risk of AI is that algorithms can create and reinforce bias. In medicine, findings from the use of AI technologies in the context of pulse oximeters, x-rays, or the determination of care based on proxies that reflect underlying inequalities have shown that the risk of bias is particularly pronounced for racially minoritized people." Amelia Fiske et al., "Weighing the Benefits and Risks of Collecting Race and Ethnicity Data in Clinical Settings for Medical Artificial Intelligence," *The Lancet* 7 (March 2025), https://www.thelancet.com/action/showPdf?pii=S2589-7500%2825%2900003-2.

15 Natalia Norori et al., "Addressing Bias in Big Data and AI for Health Care: A Call for Open Science," *Patterns* 2, no. 10 (October 8, 2021), https://pmc.ncbi.nlm.nih.gov/articles/PMC8515002/. Reva Schwartz et al., "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," U.S. Department of Commerce, National Institute of Standards and Technology, March 2022, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.

medical AI, health care worker bias in treating minority patients has been a matter of long-standing study. Additionally, traditional non-AI software, while not employing training data, can also manifest performance differences related to demographics. For example, a 2019 study found evidence of racial bias within a traditional software algorithm that used predicted health care costs as a proxy for the need for extra patient care and did not account for the factor that "unequal access to care means that we spend less money caring for Black patients than for White patients."[16]

The question of AI bias is valid in and of itself, but it is enmeshed within a larger assumption that has its own problems: adequate representation of a population segment within training data produces AI device effectiveness within that segment. As discussed later in this paper, the reasons why this assumption is flawed illuminate core challenges to AI generalization and provide a compass for improving patient safety with respect to generalization uncertainty.

AI TRAINING DATA AND GENERALIZATION

The majority of FDA-approved AI-enabled medical devices are used for medical image analysis.[17] Additional FDA-cleared devices include clinical decision support (CDS) tools that analyze patient-specific data and output specific recommendations upon which clinicians may rely. Additionally, there are CDS algorithms that are part of electronic health records (EHR) and regulated by the Office of the National Coordinator (ONC)/Assistant Secretary for Technology Policy (ASTP). In the interest of focus and avoiding generalizations too broad to be actionable, this paper limits itself to AI-enabled medical devices involved in medical image analysis and defers discussion of CDS to a future publication.

Though there are a variety of different programming architectures, AI-enabled medical devices often use artificial neural networks (ANN). Convolutional neural networks (CNN), which are a subset of ANN, specialize in visual data analysis and are popular for processing medical images used as device inputs. A neural network's processing of such inputs is ultimately expressed as a disease prediction, a categorization (such as a diagnosis), a recommendation, or some other clinical determination. Whatever the type, these determinations are generically described as device outputs.

The algorithms within a neural network's processing tend to be probabilistic. In other words, they calculate the likelihood of a particular clinical determination rather than employing static

16 Ziad Obermeyer et al., "Dissecting racial bias in an algorithm used to manage the health of populations," Science, October 25, 2019, https://www.science.org/doi/10.1126/science.aax2342

17 Rohan Singh et al., "How AI Is Used in FDA-Authorized Medical Devices: A Taxonomy Across 1,016 Authorizations," *NPJ Digital Medicine* 8, no. 308 (July 1, 2025), https://www.nature.com/articles/s41746-025-01800-1.

rules that make the determination only when a set of conditions is satisfied. For example, the Sybil AI device evaluates a single low-dose CT scan (a type of specialized X-ray image) of the lungs and uses deep learning to[18] predict a patient's risk of developing lung cancer up to six years in the future.[19] This probabilistic capacity can be life-saving because, in the case of various cancers, later detection results in a higher mortality rate.[20]

AI's probabilistic nature is powerful in health care because it is well-suited for complex medical data that may be ambiguous, incomplete, or contain subtle patterns that are not readily detected by physician review. However, AI devices need considerable amounts of information[21] to:

- improve device accuracy and maximize the range of possibilities and complexity that the device can successfully accommodate,[22]
- reduce the risk of model underfitting,
- increase training on infrequent — though medically important — edge case scenarios, and
- quantify output uncertainty.

The information used to achieve these objectives is known as training data. The type of training data varies by device category and includes (but is not limited to) x-rays, chest scans, and mammograms.[23] The visual information from such images is entered into the neural network's input layer (see Figure 1) as numeric representations of the image's pixels. This data is progressively processed within the neurons between the input and output layers. The final layer, the output layer, indicates the result produced through the calculations of the hidden layers. While the output may be textually, inside the device the output is numeric, the importance of which will be illuminated in the explanation of device training.

The most meaningful dissimilarity between traditional software development and ANN development is how parameters inside the software are set. A parameter is a variable (i.e., a

18 *Deep learning* refers to analysis by an ANN that has multiple layers of neurons between the input layer and the output layer.

19 Peter G. Mikhael et al., "Sybil: A Validated Deep Learning Model to Predict Future Lung Cancer Risk From a Single Low-Dose Chest Computed Tomography," Journal of Clinical Oncology 41, no. 12 (January 12, 2023), https://ascopubs.org/doi/10.1200/JCO.22.01345.

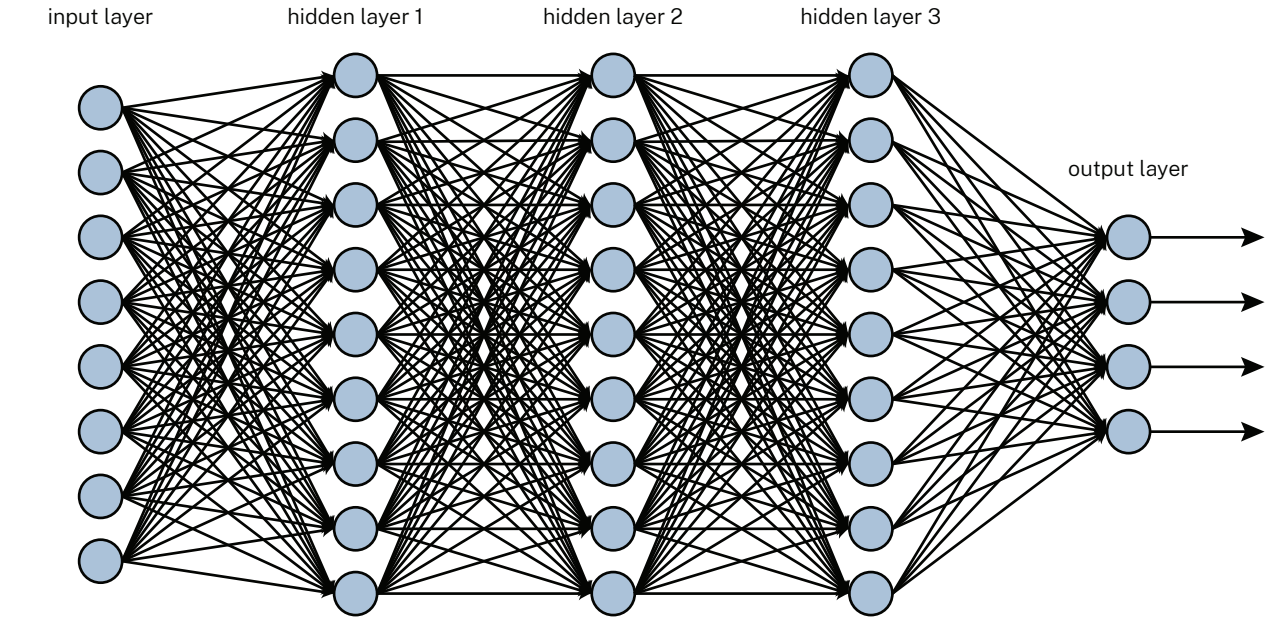
20 The American Lung Association has noted that "43% of cases are not caught until a late stage when the survival rate is only 10%." American Lung Association, "Lung Cancer Key Findings," February 5, 2026, https://www.lung.org/research/state-of-lung-cancer/key-findings.

21 Other factors affecting the amount of information needed (i.e., training data) include the degree of noise (i.e., statistical irregularity) in the data samples, the AI programming architecture employed to process the data, and the complexity of analysis performed on the information.

22 It should be noted that larger datasets do not produce improved accuracy in all situations. For a more detailed discussion of training data size and its possible effects on AI-based health care prediction models, see Richard D. Riley et al., "Importance of Sample Size on the Quality and Utility of AI-Based Prediction Models for Healthcare," *The Lancet* 7, no. 6 (June 2025), https://www.thelancet.com/journals/landig/article/PIIS2589-7500(25)00021-4/fulltext.

23 Some AI-enabled medical devices unrelated to image analysis include devices that analyze ECGs, EEGs, and glucose levels. See Singh, "How AI Is Used."

Figure 1: An Artificial Neural Network Diagram



SOURCE: National Energy Research Scientific Computing Center document Peter Harrington and Steven Farrell, "Scientific Deep Learning on Perlmutter," National Energy Research Scientific Computing Center, January 7, 2022, p. 3, https://www.nersc.gov/assets/Uploads/Deep-Learning-on-Perlmutter.pdf.

NOTE: Adapted from the original by National Energy Research Scientific Computing Center.

value such as a number or a string of letters) used to control software functions and performance. For example, in a point-of-sale system that processes credit card payments for goods, there is a parameter that sets the sales tax rate for the end user's state. In such traditional software, the parameters are manually entered into the code by a programmer. In neural-network medical devices,[24] parameters are set by algorithms interacting with training data. The parameters in such devices may number in the hundreds of thousands (or more).[25] The process of setting parameters is iterative and is described as "training" the AI model within the device. The word *model* refers to the totality of the system's algorithms and their respective parameterizations.

For each instance of training data (e.g., a single mammogram), algorithms evaluate the difference between the correct clinical determination (i.e., the output) that was expected for that one instance of training data and the observed determination that the device produced. This difference is expressed numerically, and algorithms then adjust the parameters to

24 This discussion contemplates AI-enabled clinical tools that the FDA classifies as medical devices as opposed to tools such as billing analyses or appointment-setting chatbots.

25 The number of parameters is largely determined by the number of neurons within the ANN. Within the ANN, these parameters are "biases" that affect the activation sensitivity of individual neurons and "weights" that increase or decrease the value passed from one neuron to another. Note that biases in this context are numeric settings within intraneural calculations and, in and of themselves, have no direct association with the concept of demographics as it is typically used.

minimize the difference between the expected and observed outputs. This process is repeated for every individual instance of training data, and the entire collection of training data may be processed multiple times[26] as the device is being trained.

Two key concepts in this process are backpropagation and gradient descent (those uninterested in this process discussion may skip ahead to Section II: Responses to Generalization Uncertainty and Their Limitations). Both concepts are rooted in the minimization of differences between the expected and observed outputs for the training data inputs. The differences between expected and observed outputs are measured numerically,[27] and a major challenge in reducing these differences — thereby increasing the accuracy of outputs — is avoiding overfitting the model to the training data and, consequently, undermining the model's ability to successfully generalize its functionality to new patient data encountered in the real world.

The backpropagation algorithm is a means to calculate the gradient (i.e., the direction and rate of change) of the parameters' contribution to the observed output's deviation[28] from the expected output. Backpropagation is performed for each neural network layer (see Figure 1), moving in reverse from the output layer toward the input layer. Once complete, a gradient descent algorithm[29] then uses the backpropagation's calculation to adjust the model's parameters in the correct direction (i.e., the ascent or descent of the gradient) and scope (i.e., how large a numeric change) so that the difference between accurate and actual output is minimized. In other words, it moves in the opposite direction of the gradient and modifies the parameters so that the observed output will more closely resemble the expected output. The parameter updates resulting from these techniques do not guarantee that the device can generalize successfully in every circumstance. Instead, backpropagation and gradient descent will have *optimized*[30] the model's parameters for the training data, and the testing data's validation of that parameterization is the basis for assuming that the device can generalize.

26 Parameterization is constrained by the "learning rate." The learning rate limits the magnitude of parameter adjustments for each training data instance and, as a consequence, also affects how long it takes to minimize the differences between the expected and observed outputs. One of the benefits of a learning rate is that it protects the parameters from being overcorrected by any one instance of training data.

27 The difference between expected and observed outputs for a single training data instance is measured through a loss function, while the average difference across the entire training data is measured through a cost function.

28 This deviation — the mathematical difference between expected and observed outputs — is known as a loss function.

29 There are multiple types of gradient descent algorithms, such as stochastic and momentum.

30 Device manufacturers might avoid thoroughly optimize parameters — and risk overfitting — through measures such as "dropout," where individual neurons are randomly disengaged during training so that their parameters are not adjusted for that training data instance.

RESPONSES TO GENERALIZATION UNCERTAINTY AND THEIR LIMITATIONS

Generalization uncertainty is a major concern in AI health care because a generalization problem can result in patient harm and related litigation. While these possibilities are hardly unique to AI medical devices,[31] they can negatively affect the technology's adoption. Slow adoption, for its part, can endanger the nation's leadership in medical AI and the public health advantages this leadership can bring.

Lagging adoption impedes AI leadership because valuable knowledge is gained from the technology's deployments, that is to say, implementations at health facilities. Among the reasons behind these learnings is that AI may not be "plug and play" like traditional software. In these cases, deployments generate valuable insights when the effects of variables (such as clinical protocols) are evaluated for their influence on AI device performance.[32] Moreover, aside from the real-world validation of AI functionality, deployments can produce useful data for future AI enhancements based on user interactions with AI devices.[33] For example, one study notes that "choices related to user interface design can shape interactions between humans and AI models, introducing new cognitive biases into clinical decision-making."[34] Hence, reduced usage arising from generalization uncertainty would reduce downstream learning that could, in turn, slow the pace of AI upgrades and deployment guidance improvements.

A related concern is that health care AI will become the domain of affluent health systems, with much lower adoption by lower-resourced providers, such as rural health systems and urban hospitals. Further, when deployed in lower-resourced systems, limited in-house AI expertise may lead to safety events that might not be observed in the more resourced systems. Costly correctives for generalization uncertainty inflate the total cost of AI ownership and worsen the "digital divide" between richer health systems that can afford AI devices with mitigations for generalization uncertainty and less-resourced systems that cannot.

31 Patient harm from AI, unfortunately, is often discussed in abstraction from patient harm that can occur from non-AI medical devices, human error, and other non-AI care.

32 "AI system performance can be influenced by changes in clinical practice, patient demographics, data inputs, health care infrastructure, among other factors. Such changes, commonly referred to as data drift (or concept drift, or model drift), may lead to performance degradation, bias, or reduced reliability. Additional factors such as user behavior, workflow integration, and changes to clinical guidelines may also impact system behavior in practice." FDA, "Request for Public Comment: Measuring and Evaluating Artificial Intelligence-Enabled Medical Device Performance in the Real-World," September 30, 2025, https://www.fda.gov/medical-devices/digital-health-center-excellence/request-public-comment-measuring-and-evaluating-artificial-intelligence-enabled-medical-device.

33 Jacob T. Rosenthal et al., "Rethinking Clinical Trials for Medical AI with Dynamic Deployments of Adaptive Systems," *NPJ Digital Medicine* 8, no. 252 (May 6, 2025), https://pmc.ncbi.nlm.nih.gov/articles/PMC12056174/.

34 Rosenthal et al., "Rethinking Clinical Trials."

There are a variety of broad responses to generalization uncertainty. Among the options[35] are:

- third-party device and algorithm certification,
- training data assessment, and
- physician evaluation of training data suitability.

While each of these approaches offers partial solutions, they share important limitations — most notably, limited personalization and compatibility with future adaptive AI systems. These shortcomings point to the need for another alternative.

Third-Party Device Certification

One approach to generalization uncertainty is to have an objective and independent party certify that an AI-enabled medical device will likely generalize in the real world, including for specific patient subgroups.[36] Certification would entail experimental verification of device functionality beyond the manufacturer's own testing data and would be performed independently of a specific implementation at a health system.[37] Additionally, verification criteria could be standardized for those devices performing the same function, which would make the meaning of certification clear for that device category.

This straightforward proposal must contend with several challenges. First, certification assumes that the independent party has access to the specific RWD needed for each category of AI-enabled medical device as well as the breadth of RWD needed to confirm generalization. The latter suggests the need for considerable RWD resources, because certification is being performed for the overall population with all its inherent diversity. Second, the approach lacks personalization — that is to say, the framing of a device's reliability relative to an individual patient's medical imaging or other medical information.

Third, the RWD must not only cover diverse use cases — like different patient illnesses and health states — but also image differences arising from different models and versions of the same imaging equipment. Different machines can produce different colors, contrast, and resolution for the same patient. These factors, as variables affecting device parameterization, could affect the AI devices' image interpretation. Fourth, in addition to the need for domain-specific data, domain-specific clinical knowledge is required for labeling the correct RWD

35 The discussion here is on general strategies rather than any specific commercial or professional association embodying some or all aspects of an individual strategy.

36 See John Halamka's brief discussion of public and private assurance labs. John Halamka and Paul Cerrato, "'Like Herding Cats': Making a Case for AI Assurance Labs," HealthsystemCIO.com, April 10, 2024, https://healthsystemcio.com/2024/04/10/like-herding-cats-making-a-case-for-ai-assurance-labs/.

37 If device certification were, instead, specific to each health system implementing the device and not generic for the market as a whole, there would be the prospect of using a health system's own data as part of the certification process. This approach would reduce costs for individual health systems seeking a particular device certification but increase the aggregate cost of device certification.

interpretation. Correct RWD interpretations must be produced before expected data interpretation can be compared with the device's observed outputs. Certification is further burdened by the certifier's obligation to keep current with the continual advancements across all AI device categories.[38]

Given the costs of data acquisition and expert labor, the price of certification may be too expensive for less-resourced health systems such as those located in rural areas. Additionally, certification assumes a particular historical snapshot, which would limit the certification's value when the FDA begins approving adaptive AI medical devices someday in the future. *Adaptive AI* refers to AI models built to improve by adjusting parameters through continuous learning after deployment.[39] The problem with a certification of a device during a single historical review extends beyond adaptive AI to locked algorithm devices that receive parameter updates as part of FDA-approved software updates.

Training Data Assessment

A more modest approach to the uncertainty problem is an assessment limited to a device's training data.[40] As with device certification, a training data assessment could be performed by an independent company or consultant. Characteristics of the device's dataset, such as size and demographic representativeness, could function as measures of its training data quality and suggest the device's potential to generalize.[41] Representativeness, in particular, is

> concerned with the extent to which the dataset represents the targeted population (such as patients) for which the application is intended. Whether the population of the dataset covers a sufficient range in terms of age, sex, race or other background information is the topic of the subdimension variety in demographics contained within the dimension variety. This dimension also contains the subdimension variety of data sources concerned with questions such as: Does the data originate from a single site? Were the measurements done with devices from the same or different manufacturers? Appropriately investigating such questions can provide a strong indication for the applicability and generalisability of the [machine learning] application in different environments.[42]

38 Jesse M. Ehrenfeld and Keith F. Woeltje, "The Challenges of Establishing Assurance Labs for Health Artificial Intelligence (AI)," *Journal of Medical Systems* 48, no. 110 (December 5, 2024), https://link.springer.com/article/10.1007/s10916-024-02127-2.

39 The FDA has historically approved locked algorithms that do not change post-approval.

40 See the METRIC-framework proposal in Daniel Schwabe et al., "The METRIC-Framework for Assessing Data Quality for Trustworthy AI in Medicine: A Systematic Review," *NPJ Digital Medicine* 7, no. 203 (August 3, 2024), https://www.nature.com/articles/s41746-024-01196-4

41 Schwabe et al., "The METRIC-Framework."

42 Schwabe et al., "The METRIC-Framework."

As an alternative to certification, training data assessment eliminates the need for RWD collection and the associated data labeling. This, by extension, reduces some of the cost to mitigate generalization uncertainty.

Although this approach's logic is intuitive, it does not escape several of the difficulties already noted for certification. First and foremost, while desirable, a demographically representative dataset does not guarantee that the device will generalize to the dataset's represented population segments. If, for example, the training data is demographically diverse but the visual characteristics within individual data samples are very similar to one another, then outlier patients (whose medical images significantly deviate from these characteristics) may be at higher risk for device non-generalization despite belonging to adequately represented demographic segments.

This predicament reflects the nature of how a neural network extracts features from an image.[43] This process starts with the image's pixels expressed as a matrix of numeric values. If, for example, the image was 500 x 500 pixels, then the matrix would have 250,000 numeric values representing the sum of all pixels (500 rows multiplied by 500 columns). These numeric values represent each pixel's color. Grayscale images have a single matrix of numbers for analysis, while color images have multiple related matrices for each pixel's color components (e.g., red, green, and blue values). A separate and very small matrix, known as a kernel or filter, moves across an image matrix using small, orderly movements. The kernel has its own values within its matrix, and these values, when applied to the specific image patch, help extract simple features (see Figure 2). At the most basic level, these extracted features may concern edges, patterns, textures, certain color values, and so on.[44] When the extracted features are analyzed in combination with one another, they can detect more complex structures and conditions within the image.

As referenced in the prior certification discussion, medical image attributes extend beyond a patient's anatomical characteristics. Differing imaging equipment, or generations of the same machine, can affect the resulting medical image and influence parameterization on the development side and generalization on the deployment side. There is also the matter of technician effects on medical images. For example, a sonographer performing an echocardiogram generates a digital record of heart behavior (e.g., wall movement, valve regurgitation, etc.). The sonographer's choices regarding the plane of the ultrasound beam during the echocardiogram affect important measurements such as blood flow velocity and heart wall thickness.

43 CNNs are a popular means for feature extraction from images.

44 This convolutional process is more elaborate and complex in actual practice.

Figure 2: Simplified Representation of an Image Matrix and a Kernel Used to Detect a Diagonal Pattern

Image with kernel in upper left position

9 *-1	0 *-1	7 *2	4	5	1	8	3	3
4 *-1	4 *2	8 *-1	7	0	1	3	1	4
2 *2	3 *-1	2 *-1	1	0	1	1	5	9
5	4	7	8	9	0	7	4	7
0	5	8	2	4	4	8	7	3
0	1	7	8	2	3	2	1	1
0	9	3	1	5	4	7	8	2
0	6	1	8	0	5	8	2	5
5	7	2	2	0	1	7	8	5
4	4	5	8	6	1	7	8	9
3	5	5	7	6	2	8	2	4
8	3	3	9	0	7	7	8	2
3	1	4	4	4	8	3	1	5
1	5	9	2	3	2	1	8	0
7	4	7	5	4	7	2	2	0
8	7	3	0	5	8	5	8	6
9	0	7	0	1	7	5	7	6

Kernel

-1	-1	2
-1	2	-1
2	-1	-1

Finally, not being an ongoing activity, a training data assessment is also ill-suited for future adaptive AI devices that change their parameters over time through continuous learning (though an analogous issue applies to nonadaptive systems where the device parameters are modified through a software update).

Physician Evaluation of Training Data Suitability

A third possible response to generalization uncertainty is physician evaluation of a device's training data before the device is used with an individual patient. A physician, using a description provided by the manufacturer, would evaluate the degree to which a device's training data encompasses a patient's demographic attributes and then would decide if the device is appropriate for use with that patient.

Physician evaluation carries with it all the challenges of training data assessment while introducing an individual patient's demographic characteristics in relation to the training data. This approach assumes a physician will consistently review a device's training data description for each patient, which is open to debate given the frequency of physician neglect of drug

labels.[45] This approach also assumes the patient will undergo medical imaging at the physician's health system where the physician would presumably have access to the device's training data description (though the manufacturer may offer a training data description publicly in the form of an "applied model card").[46]

A related concern is how much information a manufacturer will publicly disclose about a device's training data.[47] Training data, as the foundation of device performance, represents a valuable intellectual property asset. Consequently, a manufacturer may craft its data description around competitive considerations. Additionally, some manufacturers may withhold training data descriptions because critics and competitors might use them to shame the manufacturer publicly about insufficient demographic diversity.

A BETTER RESPONSE TO GENERALIZATION UNCERTAINTY

The shortcomings in generalization uncertainty responses provide valuable context for constructing an alternative. Ideally, an alternative would:

- personalize generalization uncertainty from the perspective of individual patients;
- reduce or eliminate the need for highly specialized and expensive data science consultants;
- avoid dependence on RWD, whether collected or purchased;
- accommodate differences among images arising from the use of different machines or imaging technicians;
- empower physicians to make more informed patient decisions about AI-enabled medical device use without requiring manual analysis;
- be compatible with either static or adaptive datasets (even though current FDA-approved medical devices have static datasets); and
- preserve training data confidentiality and its control by the device manufacturer.

45 Mari Serebrov, "If No One Reads It, What's the Purpose of a Drug Label?," *BioWorld*, March 29, 2018, https://www.bioworld.com/blogs/1-bioworld-perspectives/post/247-if-no-one-reads-it-what-s-the-purpose-of-a-drug-label-.

46 See the development data characterization section of the CHAI's applied model card at https://mc.chai.org/v0.1/documentation.pdf. The model cards themselves are intended to be housed in a centralized registry open to both public and clinician access. Alexis Kayser, "Health Care's First AI Registry Is Coming Soon," *Newsweek*, February 28, 2025, https://www.newsweek.com/nw-ai/health-care-artificial-intelligence-ai-applied-model-cards-registry-chai-2037434.

47 For comparison, a study of 903 AI-enabled medical devices found for devices that reported clinical performance studies that "less than one-third of the clinical evaluations provided sex-specific data, and only one-fourth addressed age-related subgroups." Clinical performance data in the study was obtained through the FDA website and other online resources. Windecker, "Generalizability."

This section proposes a generalization uncertainty response described as Digital Similarity Analysis (DSA). Unlike existing approaches, the DSA proposal does not rely on broad population-level validation or costly third-party certification. Instead, it provides a patient-specific assessment of whether an AI device is likely to generalize effectively. The final section of the paper reviews some implications of generalization uncertainty from a policy perspective.

Rethinking the Training Data Question

A central premise related to generalization uncertainty is that training data, as the basis for device parameterization, greatly influences generalization results. As explained earlier, the difference between the expected output for each training data instance and the observed output is minimized by iteratively modifying the device's parameters, thus improving its accuracy. When training is fully completed, the device is validated on a smaller set of testing data.

To avoid underfitting (and compromising generalization), manufacturers employ an assortment of tactics. One of the most basic is training the device on a large and informationally diverse dataset. Large datasets, it is hoped, will provide sufficient pattern variations that, once reflected in the parameter settings, will enable the device to accommodate more RWD scenarios than would be the case with a narrow dataset. In the case of AI-enabled medical devices, the dataset is typically patient medical images whose correct interpretations (i.e., the expected output) are known prior to the training of the AI system.

However, even when training data is varied and demographically representative, generalization is not guaranteed because device performance depends on underlying image characteristics and their extraction. If a medical device's training data is highly dissimilar to a patient's own medical image, there are reasonable questions[48] about whether this device will generalize for this individual. These doubts arise because the device's parameters are adjusted by the visual characteristics of the training data,[49] and different visual characteristics can result in different parameter modifications. Consequently, reviewing a written demographic summary of training data does not diminish generalization uncertainty, nor does a prior validation test that had not anticipated this patient's deviation from the training dataset. In both cases, there remains an unconsidered — and potentially dangerous — mismatch between the characteristics of the patient image and those represented in the device's training data.

48 Questions about the proper function of an AI device, it should be noted, do not necessarily mean an AI device will fail to generalize, but they do mean that there is a reasonable concern that this may be the case.

49 Kernels (otherwise known as filters) in a CNN interact with these characteristics to detect those features that are related to the device's purpose (diagnosis, prediction, classification, etc.).

Instead of accepting this state of affairs and its accompanying AI safety and adoption concerns, we can mitigate some degree of risk through "selective deployment" informed by the similarity between the patient's medical image and the device's training data.[50]

Selective Deployment

In their paper "The Selective Deployment of AI in Healthcare,"[51] Robert Vandersluis and Julian Savulescu consider the underrepresentation of groups within training data. The authors use the phrase *selective deployment* to describe the use of AI with those populations for whom the AI performs well. The first case study in support of this approach was a prognostic[52] algorithm for breast cancer, and the second, a skin disease algorithm. In the case of the breast cancer algorithm, the training data was exclusively female given data collection challenges related to women suffering from breast cancer much more frequently than men.[53] In the case of the skin disease algorithm, there was a different disparity in training data representation. Light-skinned people were overrepresented compared to the population as a whole, as melanoma is much more prevalent among people with lighter skin.[54]

Vandersluis and Savulescu argue that instead of deploying the algorithm for general use with any patient, female or male, with breast cancer, it is better to use the algorithm with "patient subgroups where the model performs well, while withholding the algorithm from those patient subgroups for whom the model is expected to perform poorly (or unpredictably)."[55] They state:

> We believe that the ethical tensions arising from the delayed and expedited deployment options are sometimes best resolved through a selective deployment approach; this approach is not ideal, but instead — with appropriate regulation — represents the best way to balance harm prevention, utility and fairness considerations.[56]

Selective deployment for the skin disease algorithm, on the other hand, was not as clear for the authors, as this algorithm's higher incidence of false positives among darker-skinned people would lead to higher rates of specialist care for that population.

50 While this discussion focuses on images, the basic issue of similarity between patient data and training data can have applications beyond visual information.

51 Robert Vandersluis and Julian Savulescu, "The Selective Deployment of AI in Healthcare," *Bioethics* 38, no. 5 (February 16, 2024), https://onlinelibrary.wiley.com/doi/10.1111/bioe.13281.

52 A prognostic algorithm predicts how a disease will progress.

53 Vandersluis and Savulescu, "The Selective Deployment of AI in Healthcare."

54 "For every 30 light-skinned patients that suffer from melanoma, only one dark-skinned patient will be in the same position." Vandersluis and Savulescu, "The Selective Deployment of AI in Healthcare."

55 Vandersluis and Savulescu, "The Selective Deployment of AI in Healthcare."

56 Vandersluis and Savulescu, "The Selective Deployment of AI in Healthcare."

The authors' advocacy of selective deployment does not preclude bias mitigation efforts, such as increasing population diversity in data collection and clinical research.[57] Rather, the approach seeks to balance lowering patient harm risk in some groups with the promise of medical benefits for others. Selective deployment seeks to reduce both direct and indirect harm by approving AI tool access for populations where indications suggest that use is appropriate, while discouraging use where it is not.

However, if a selective deployment decision is limited to the training data's demographic composition, in some circumstances AI device benefits may be needlessly withheld from groups despite the intention to prevent patient harm. The underrepresentation of a patient's demographics in training data, though undesirable, does not confirm that an AI device will fail to generalize for that patient — just as adequate representation of a patient's demographics does not guarantee successful generalization. *A preferable framework would personalize a selective deployment decision based on characteristics related to an individual patient.* This premise is the foundation of the Digital Similarity Analysis proposal.

Digital Similarity Analysis

DSA proposes to evaluate the resemblance (similarity/dissimilarity) between an individual patient's medical image[58] and the domain of medical images comprising a device's training data and testing[59] data. The purpose of DSA is to determine if a patient's medical image is an outlier compared to this data *prior to the use of that device with the patient image*. The physician, when alerted by DSA that a patient's image is significantly dissimilar to the AI device's training and testing data, can make a selective deployment decision to (a) forgo device use, (b) require supplemental validation of the medical device's output for the patient, or (c) use the device but treat any clinical determination produced by the device with lower confidence. It is important to note that an outlier judgment by the DSA approach does not predict a generalization failure; rather, it highlights when an image's attributes were not well represented in either the AI model's parameterization (training) or validation (testing).

The DSA hypothesis begins with the comparison of an individual patient's medical image to the training and testing data used to develop an AI medical device. This comparison does not seek identical matches, because an AI medical device is not an index of images. CNN

57 See Box 1 within Vandersluis and Savulescu, "The Selective Deployment of AI in Healthcare."

58 *Image*, in the context of the DSA, refers equally to a static image or a sequence of images within a video or a data series such as an EKG. Additionally, the same approach could be applied to audio files.

59 The inclusion of testing data alongside training data will be explained later in this section.

techniques,[60] such as feature extraction and pooling,[61] produce an abstraction of a patient medical image that is then processed through the remaining layers of the neural network. Accordingly, the comparison determines whether the patient image belongs to the same domain (i.e., family of characteristics) as the images used for device training and testing.

Determining visual similarity, while intuitive for humans, is a sophisticated operation for computers. A host of variables complicate the process of distinguishing an image feature from its larger context. Figures 3 and 4, while simplistic, illuminate some basic contours of the challenge. Starting with Figure 3, two identical representations of the letter A are displayed side-by-side using the same color and pixel pattern. However, in the second, the letter A is shadowed and requires a computer to distinguish the letter shape from both the background's lighter contrast pixels and the darker contrast black pixels of the shadow.

Figure 4 illustrates how, even in the absence of contrast complexities (for letters sharing the same identity, color, and typographical pattern), the additional factors of size and rotation present interpretation issues that a computer must navigate. Pattern issues such as line width, edge orientation, and the distance between the character's crossbar and apex differ among each letter A.

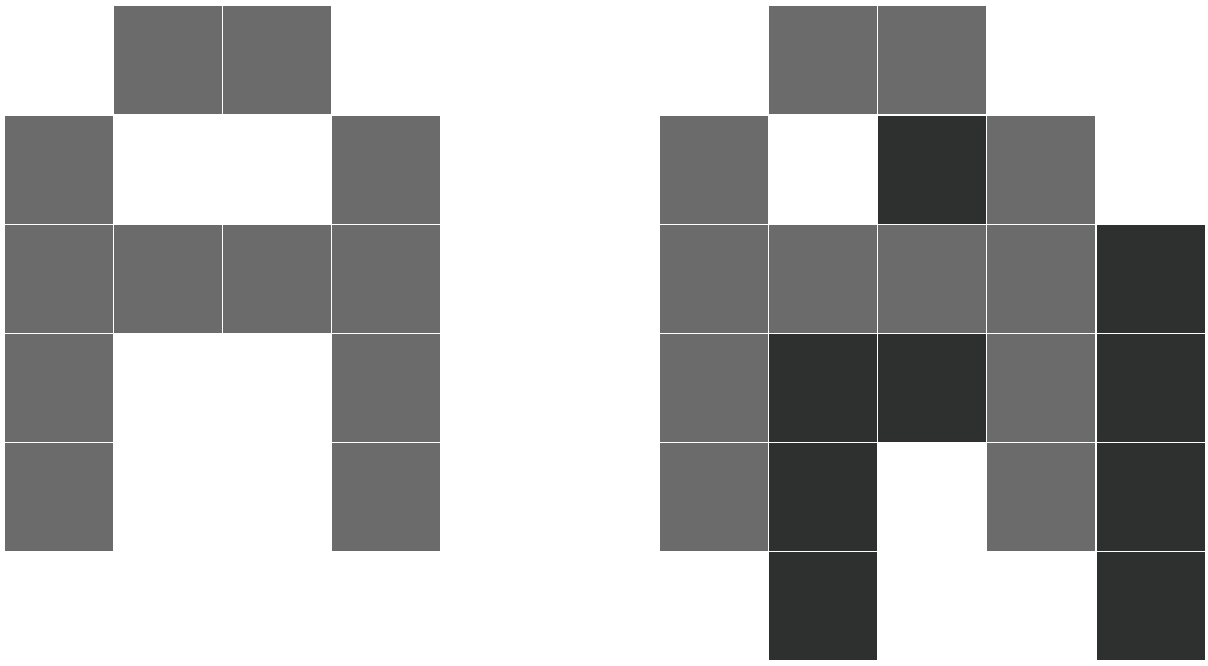
Fortunately, computer science has produced a wealth of image analysis algorithms that can be employed for DSA alongside novel algorithms for unaddressed tasks. Some existing image analysis algorithms relevant to this activity are:

- Structural Similarity Index Measure (SSIM): Originally developed for the quality assessment of an image relative to its source (such as after lossy image compression), SSIM can assess similarity by comparing attributes such as contrast and luminance (i.e., brightness).
- Cosine Similarity (CS): CS calculates the similarity between two vectors (e.g., linear patterns) detected within an image and may be used within the larger process of analyzing the structures within an image regardless of scale and rotation.
- Scale-Invariant Feature Transform (SIFT): SIFT identifies features in a source image and evaluates comparison images for the presence of these features regardless of size and rotation.

60 A discussion of CNN data processing can be reviewed in Keiron O'Shea and Ryan Nash, "An Introduction to Convolutional Neural Networks," Arxiv.org, December 2, 2015, https://arxiv.org/pdf/1511.08458.

61 Pooling is an operation within a CNN that reduces the spatial dimensions of the features from an image processed by an AI-enabled medical device. Pooling reduces subsequent computational requirements while preserving the image characteristics necessary to produce the device's output.

Figure 3: Pixel Differences Between Identical Letters Unshadowed and Shadowed



- Oriented FAST and Rotated BRIEF (ORB): ORB, like SIFT, identifies image features but ORB is optimized for computational speed.
- Fréchet Radiomic Distance (FRD): FRD is an algorithm created for radiology images and performs a variety of tasks including detecting the main features that differ between image sets and identifying out-of-domain characteristics.[62]
- Speeded-Up Robust Features (SURF): SURF[63] employs a fast computation process of feature extraction, description, and matching.[64]

Discriminators within generative adversarial networks (GANs) may also prove helpful for DSA development. A GAN uses a software "discriminator" to determine whether an AI-generated image is sufficiently similar to confirmed examples within the image class being mimicked

62 Nicholas Konz et al., "Fréchet Radiomic Distance (FRD): A Versatile Metric for Comparing Medical Imaging Datasets," *Medical Image Analysis*, June 6, 2025, https://arxiv.org/abs/2412.01496.

63 Herbert Bay et al., "SURF: Speeded Up Robust Features," ETH Zurich/K. U. Leuven, September 10, 2008, https://web.archive.org/web/20220120220703/ftp://ftp.vision.ee.ethz.ch/publications/articles/eth_biwi_00517.pdf.

64 MATLAB, "Feature Extraction Using SURF," https://www.mathworks.com/help/gpucoder/ug/feature-extraction-using-surf.html.

Figure 4: Identical Characters of Different Sizes and Orientation

(i.e., generated by the GAN for an end user). Beyond GANs, there are other pretrained networks commercially marketed for image analysis.[65]

While not solving every challenge of image similarity analysis in AI-enabled medical devices, existing resources provide a strong foundation from which the DSA hypothesis can be explored. Moreover, preexisting algorithms for image analysis have been validated in tasks such as facial recognition, categorizing objects within a photo, retrieving similar images from repositories, recognizing duplicate images, and detecting image degradation after file compression. The algorithm list itself does not exhaust the many options that can assist in determining similarity between an individual patient's medical image and a device's training and testing images. DSA's similarity judgment should be a composite score involving multiple dimensions:

- The patient image's similarity to the total collection of training data
- The percentage of files within the training data that have a high similarity to the patient image
- The patient image's similarity to the total collection of testing data used for device validation

65 ResNet is a popular example of a pretrained image analysis network. Wanni Xu et al., "ResNet and Its Application to Medical Image Processing: Research Progress and Challenges," *Computer Methods and Programs in Biomedicine* 240 (October 2023), https://www.sciencedirect.com/science/article/abs/pii/S0169260723003255.

- The highest similarity score between the patient image and any one instance of testing data

Testing-data similarity should be given even more weight within similarity scoring than training-data similarity, because the testing data has been used to validate the performance of the AI medical device. Training data, while used to adjust the model's parameters, has some opaque dependencies. When backpropagation is performed to minimize the difference between a training example's expected and observed outputs, the magnitude of the parameter adjustments is constrained by the learning rate as well as the number of training epochs.[66] Neither variable is known outside the manufacturer who developed the device.

DSA avoids replicating each medical device's custom set of kernels so that its feature extraction is identical to that of the medical device. Such replication would not only require a device manufacturer to divulge highly confidential trade secrets, but it would also make the DSA process costly and highly customized. Instead, DSA recognizes that a device's selection and representation of features is derivative of more basic visual attributes present in the raw image input prior to feature extraction. DSA would operate at an intermediate level of image comparison that, while less detailed than the highly specialized kernels of the medical device, would still discriminate significant features upon which meaningful similarities and dissimilarities can be established. The operationalization of this comparison, however, requires empirical testing to establish dissimilarity thresholds that justify selective deployment suggestions.

DSA does not require training and testing data to be either publicly disclosed or publicly rated. In theory, the DSA process could be implemented either locally at the device manufacturer[67] or remotely within a private and HIPAA-compliant cloud environment provided by a DSA vendor. In either scenario, DSA would be made accessible to medical facilities that own DSA-supported devices. These facilities would securely upload patient images via the internet for comparative evaluation. The manufacturers themselves would have an incentive to provide DSA to providers, because DSA:

- supplies patient-level evaluation that lowers a medical device's perceived generalization risks in the eyes of health systems considering its purchase,
- avoids certification or assessment consulting expenses that could significantly increase a device's total cost of ownership,

66 An epoch is a full cycle of AI model training where every file in the training data has been used to update parameters. The successful training of an AI model may require multiple epochs.

67 In this scenario, the manufacturer would host the DSA software in a HIPAA-compliant environment and run the analyses itself for providers using its device.

- increases transparency around training images without compromising their confidentiality, and
- demonstrates a good faith effort to mitigate patient safety issues associated with generalization uncertainty, not only for groups underrepresented in the device's training data but for patients overall.

An important dimension of the DSA proposal is the capacity for the tool to be syndicated, that is to say, offered as standardized software as opposed to a custom project for each facility that desires it. Syndication could reduce implementation costs for DSA as compared to a highly custom/consultative offering. Syndication allows development expenses to be spread out across clients and lower the marginal costs of adding new clients. Thus, syndication would ideally make the tool more financially accessible for rural health systems and under-resourced urban facilities.

Comparison of Competing Responses to Generalization Uncertainty

When contrasted against the previously discussed responses to generalization uncertainty, DSA has multiple advantages as demonstrated in Table 1.

A last matter not addressed in Table 1 are the challenges surrounding image consistency and quality from one imaging device to another. The value of a device certification and training data assessment is problematized by RWD images that differ in color, contrast, brightness, size, resolution, and other qualitative differences that are not related to the patient. Such differences could materially interfere with a medical device's generalization, as these equipment-related differences might impair a neural network's perception of the patterns it was designed to detect.[68] DSA may prove especially helpful in this context, as these equipment-related differences would be evident within the image comparison process.

POLICY IMPLICATIONS

The DSA proposal could make a valuable contribution to AI medical device safety and avoid alternatives that inadequately address the problem. Further, DSA extends the discussion of algorithmic bias from broad demographic categories to individual patient characteristics. The proposal also has the additional benefit of addressing image variation arising from differences in radiology equipment and technician techniques.

68 "Successful translation of artificial intelligence (AI) models into clinical practice, across clinical domains, is frequently hindered by the lack of image quality control. Diagnostic models are often trained on images with no denotation of image quality in the training data; this, in turn, can lead to misclassifications by these models when implemented in the clinical setting." Syed Rakin Ahmed et al., "Generalizable Deep Neural Networks for Image Quality Classification of Cervical Images," *Scientific Reports* 15, no. 6312 (February 21, 2025), https://www.nature.com/articles/s41598-025-90024-0.

Table 1: Comparison of Generalization Uncertainty Mitigations

Issue	Device Certification	Training Data Assessment	Physician Evaluation	Digital Similarity Analysis
Personalizes Generalization at the Patient-Level?	No.	No.	Yes.	Yes.
Requires a Collection of Many RWD Samples?	Yes.	No.	No.	No.
Compatible with Adaptive Training Datasets?	No. Certification is for a single point in time. However, certification could be combined with postmarket monitoring.	No. Changes to a device's training dataset would require a new assessment.	Yes, so long as the device manufacturer periodically updates the training data description.	Yes, so long as the device manufacturer periodically re-runs the feature extraction on the training data.
Automated?	Unlikely. While subtasks of certification could be automated, other labor (such as RWD labeling) would require manual effort unless a commercial prelabeled dataset was acquired.	Possibly. Assuming demographic attributes were known for every medical image, the determination of training data breadth could be automated.	No.	Yes.
Quality and Consistency of Analysis	Low-to-medium consistency, because certification is highly consultative and ends in a binary certified/ uncertified judgment.	High consistency if training data evaluation criteria are standardized.	Low consistency because of human subjectivity and lack of standardization for evaluation.	High consistency of analysis due to algorithmic approach.
Estimated Expense	Potentially high cost due to data, clinical, and AI expertise needed for certification.	Unclear due to uncertain scope of automation. Potentially medium to high cost if the data and clinical work within the assessment is manual.	Potentially low cost limited to physician time spent on evaluating training data suitability for patient based on training data summary description.	Potentially low cost, as the analysis employs a commercially syndicated tool that can be implemented with negligible marginal costs for additional AI-enabled medical devices.
Worsens Digital Divide Between Well-Resourced Health Systems and Low-Funded Rural/Urban Health Systems?	Likely. This approach could exacerbate the digital divide due to expenses of RWD acquisition and high-priced human labor involved in data labeling and device validation experiments.	Likely. This approach could exacerbate the digital divide due to the expenses of RWD acquisition as well as high-priced human labor involved in data labeling and device validation experiments.	Unlikely. This approach does not meaningfully increase the total cost of ownership for AI-enabled medical devices.	Unlikely. As it is a syndicated solution, costs are minimized by spreading them out across all solution users and avoiding expensive custom consulting.
Intellectual Property Protection	Assuming certification is limited to an empirical device validation and does not include a training data assessment, the manufacturer's training data does not need to be disclosed to the certifiers.	The device manufacturer's training data must be disclosed to a third party for assessment, but there may be an opportunity for automation that would avoid manual review of the data.	Training data is demographically summarized by the device manufacturer, and that summary is provided to a physician. Physician evaluation does not publicly disclose attributes of the dataset.	DSA feature extraction from the training and testing data does not require public disclosure and would be securely restricted from non-manufacturer access.

By improving the grounds for AI medical device adoption, DSA aligns well with the nation's larger AI health policy aspirations, such as "reduce[d] barriers to the use of AI technologies to promote their innovative application."[69] Moreover, DSA use does not preclude other complementary safety efforts. Notable in this context is targeted postmarket surveillance for AI-enabled medical devices at risk for future unpredictability,[70] as well as efforts to predict generalization.

From a policy perspective, *DSA does not need major legislative or regulatory interventions or government grants*. Freedom from these entanglements insulates the DSA proposal from the debates that new rules attract as well as accusations of government favoritism. Additionally, as a tool for evaluating medical device suitability — and not delivering the medical functionality itself — DSA does not require FDA approval or a regulatory sandbox.[71] Leveraging preexisting image analysis algorithms could reduce both the tool's development cost and time to market.

With respect to the health care market, DSA use would be voluntary, and its success is not dependent on any mandates around training data transparency,[72] sourcing,[73] or composition requirements.[74] This distinguishes DSA from alternatives that rely on mandates. The State of California, for example, passed the Generative Artificial Intelligence: Training Data Transparency Act.[75] This law requires generative AI systems to publicly disclose a variety of characteristics of the data that was used to train these systems. Washington State has proposed a similar bill.[76]

While free of government dependencies, the validation of DSA could be assisted (post-development) by a pilot project within the Veterans Health Administration (VHA). As noted in the paper "Could the VA Be the Key to Lowering the Cost of American Health Care?",[77] the

69 Exec. Order No. 13859, 84 Fed. Reg. 3967 (Feb. 11, 2019), https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence.

70 Kev Coleman and Michael Pencina, "Targeted Postmarket Surveillance: The Way Toward Responsible AI Innovation in Health Care," Paragon Health Institute, September 2025, https://paragoninstitute.org/private-health/targeted-postmarket-surveillance-the-way-toward-responsible-ai-innovation-in-health-care/.

71 A regulatory sandbox is a governance framework that establishes a defined, short-term duration (e.g., one year) that allows an AI manufacturer to demonstrate a product or service on a small scale and allow policymakers to evaluate its results as a basis for future evidence-based rulemaking.

72 See California's law AB-2013, "Generative Artificial Intelligence: Training Data Transparency," at https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2013.

73 Ibid.

74 Sean Miller, "What Are the Diversity Requirements for AI Training Data?," Luth Research, February 16, 2026, https://luthresearch.com/glossary/what-are-the-diversity-requirements-for-ai-training-data/.

75 https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2013

76 Valerie Shmigol and Ana Gonzalez, "AI Transparency: A Tale of One Task Force and Two States' Legislative Bills," Summit Law Blog, November 14, 2025, https://www.summitlaw.com/law-blog/ai-transparency-a-tale-of-one-task-force-and-two-states-legislative-bills.

77 Kev Coleman, "Could the VA Be the Key to Lowering the Cost of American Health Care?," Paragon Health Institute, July 16, 2025, https://paragoninstitute.org/wp-content/uploads/2025/07/Could-VA-Solve-Healthcare-Spending_RELEASE_V2.pdf.

VHA provides numerous assets with regard to AI. The VHA has the scale and geographic breadth to provide a source of RWD that represents diverse populations and health trends that could robustly test DSA's effectiveness. With regard to this breadth, the agency serves over 9.1 million veterans and operates 1,380 health care facilities across the United States.[78] Additionally, the VHA, being part of the Department of Veterans Affairs, comes under the agency's "Fourth Mission," a directive that includes the support of public health efforts.[79]

Lastly, should the DSA proposal prove successful, the learnings could be extended to analogous AI challenges related to the interpretation of non-image inputs (e.g., EKGs), serial images in video, or devices analyzing sound patterns (e.g., AI-enabled stethoscopes).

78 VA, "Veterans Health Administration," https://www.va.gov/health/.

79 "VA's 'Fourth Mission' is to improve the nation's preparedness for response to war, terrorism, national emergencies, and natural disasters by developing plans and taking actions to ensure continued service to Veterans, as well as to support national, state, and local emergency management, public health, safety and homeland security efforts." VHA, "VA Fourth Mission Summary," May 9, 2022, https://www.va.gov/health/coronavirus/statesupport.asp.

APPENDIX: LIST OF SELECTED AI ACRONYMS

ANN: Artificial Neural Network. An ANN is a form of machine learning that uses layers of artificial neurons (called nodes) to process data entered into the ANN. Each node within a layer is an individual software module that, as part of its collaboration on a shared ANN task, processes the data passed into it and then provides the result to neurons in the next layer of the ANN.

CNN: Convolutional Neural Network. A CNN is a type of ANN that is optimized for the analysis of visual data within an image. A CNN extracts various features within a medical image in order to perform a disease prediction, diagnosis, recommendation, or other clinical determination.

DSA: Digital Similarity Analysis. An approach to AI-enabled medical device safety that evaluates the similarity/dissimilarity of a patient medical image to the data used in a specific medical device's parameterization (training) and validation (testing) processes. If the patient image is significantly dissimilar to the device data, the clinician is alerted that the patient image represents an outlier condition and the clinician should consider (a) forgoing device use, (b) requiring supplemental validation of the medical device's output for the patient, or (c) using the device but treating any clinical determination produced by the device with lower confidence.

GAN: Generative Adversarial Network. A GAN is a form of AI where original content (text, speech, images, etc.) is produced through the interactions of two ANNs: a generator and a discriminator. The generator generates novel content, and the discriminator evaluates the content. The discriminator is adversarial because it compares the new content to examples of content in the same category that are already confirmed to be authentic. If the discriminator can distinguish the new content from the confirmed content, the generator must improve the content through subsequent iterations until the discriminator cannot reliably differentiate between the new content and the examples.

RWD: Real World Data. Patient data sourced from the medical facilities at which AI-enabled medical devices are deployed. RWD is separate from an AI device's training and testing data.

VHA: Veterans Health Administration. The VHA delivers health care services to American veterans.