npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02052-9

Evaluating transparency in AI/ML model characteristics for FDA-reviewed medical devices

Check for updates

Viraj Mehta[1], Abhinav Komanduri[2], Rishabh Singh Bhadouriya[3], Vilina Mehta[4], Michael David Johnson[4], Priyanka Shrestha[1], Margaret Nikolov[5], Bhav Jain[4], Nigam Shah[5,6] & Kevin Schulman[5,7] ✉

The rapid integration of artificial intelligence (AI) and machine learning (ML) into medical devices has underscored the need for transparency in regulatory reporting. In 2021, the U.S. Food and Drug Administration (FDA) issued Good Machine Learning Practice (GMLP) principles, but adherence in FDA-reviewed devices remains uncertain. We reviewed 1,012 summaries of safety and effectiveness (SSEDs) for AI/ML-enabled devices approved or cleared by the FDA between 1970 and December 2024. Transparency in model development and performance was assessed using a novel AI Characteristics Transparency Reporting (ACTR) score across 17 categories. The average ACTR score was 3.3 out of 17, with modest improvement by 0.88 points (95% CI, 0.54–1.23) after the 2021 guidelines. Nearly half of devices did not report a clinical study and over half did not report any performance metric. These findings highlight transparency gaps and emphasize the need for enforceable standards to ensure trust in AI/ML medical technologies.

The emergence of increasingly powerful artificial intelligence technologies in health care has driven a debate over their regulation[1–4]. Artificial intelligence (AI) and machine learning (ML) are becoming increasingly integrated into medical devices for the diagnosis and care of patients[5,6]. In health care, the Food and Drug Administration (FDA) is the regulatory agency with authority over these technologies when they meet criteria as medical devices. Evaluation of the FDA's regulatory oversight of these devices could provide important insights into this national debate.

In October 2021, the FDA established a set of 10 guiding principles for good machine learning practice (GMLP) for machine learning-enabled medical devices (MLMDs) in collaboration with Health Canada and the United Kingdom (UK) Medicines and Healthcare Products Regulatory Agency (MHRA)[7]. In one of these ten principles, they mandate that "users are provided clear, essential information," including "performance of the model for appropriate subgroups, characteristics of the data used to train and test the model, […] device modifications and updates from real-world performance monitoring"[7].

As of December 2024, the FDA has approved or cleared 1016 medical devices for marketing that use AI/ML through its 510(k), premarket approval (PMA), and De Novo pathways, with 572 devices approved or cleared since the 2021 guidelines[8]. To assess the degree to which MLMDs approved or cleared by the FDA have adhered to their own reporting principles, and whether the 2021 guidance influenced the quality of the reporting, we systematically reviewed the publicly available Summary of Safety and Effectiveness (SSED) for each of the 1016 devices.

Results

We analyzed approval summaries for 1016 total MLMDs, of which 1012 were accessible. The device approval dates spanned from 1970 to 2024, with 99.7% of device approvals after 2000 and 54.2% after 2021.

Regulatory Variables

We found that 96.4% of devices were cleared via the 510(k) pathway ($n = 976$), followed by the De novo pathway ($n = 32$, 3.2%) and premarket approval pathway ($n = 4$, 0.4%) (Table 1). Clinical specialties include radiology ($n = 769$, 76%), cardiovascular medicine ($n = 99$, 9.8%), neurology ($n = 37$, 3.7%), hematology ($n = 18$, 1.8%), gastroenterology/urology ($n = 16$, 1.6%), anesthesiology ($n = 14$, 1.4%), ophthalmology ($n = 11$, 1.1%), and others ($n = 48$, 4.7%), which primarily included clinical chemistry, general & plastic surgery, and orthopedics (Table 1).

[1]Stanford University, Department of Computer Science, Stanford, USACA. [2]University of Southern California, Department of Population and Public Health Sciences, Los Angeles, USACA. [3]University of Pittsburgh, College of General Studies, Pittsburgh, USAPA. [4]Stanford University, School of Medicine, Stanford, USACA. [5]Clinical Excellence Research Center, School of Medicine, Stanford University, Stanford, USACA. [6]Technology and Digital Solutions, Stanford Healthcare, Stanford, USACA. [7]Operations, Information and Technology (by courtesy), The Graduate School of Business, Stanford University, Stanford, USACA. ✉e-mail: Kevin.Schulman@Stanford.Edu

Table 1 | Number of AI/ML-enabled devices authorized by the Food and Drug Administration (FDA) that report on characteristics of interest in approval summaries

Characteristic	No. Reported, n (%)	ACTR Score Input?
Clearance pathway		N
510(k)	976 (96.4)	-
De novo	32 (3.2)	-
Premarket approval	4 (0.4)	-
Specialty panel		N
Radiology	769 (76)	-
Cardiovascular medicine	99 (9.8)	-
Neurology	37 (3.7)	-
Hematology	18 (1.8)	-
Gastroenterology/Urology	16 (1.6)	-
Anesthesiology	14 (1.4)	-
Ophthalmology	11 (1.1)	-
Other	48 (4.7)	-
Predetermined change control plan (PCCP)		Y
Reported	15 (1.5)	-
Not reported	997 (98.5)	-
Clinical study		Y
No clinical study conducted	475 (46.9)	-
Data collection not reported	126 (12.5)	-
Retrospective data collection	325 (32.1)	-
Prospective data collection	75 (7.4)	-
Retrospective & prospective data collection	11 (1.1)	-
Clinical study sample size	403 (39.8)	Y
Train dataset source		Y
Exact sites	18 (1.8)	-
Number of sites	8 (0.8)	-
Region of sites	33 (3.3)	-
Number & region of sites	9 (0.9)	-
Not reported	944 (93.3)	-
Train dataset size		Y
Patients only	30 (3.0)	-
Images only	49 (4.8)	-
Patients and images	16 (1.6)	-
Neither	917 (90.6)	-
Test dataset source		Y
Exact sites	36 (3.6)	-
Number of sites	48 (4.7)	-
Region of sites	68 (6.7)	-
Number & region of sites	96 (9.5%)	-
Not reported	764 (75.5)	-
Test dataset size		Y
Patients only	71 (7.0)	-
Images only	116 (11.5)	-
Patients and images	48 (4.7)	-
Neither	777 (76.8)	-
Dataset demographics		Y
Reported	240 (23.7)	-
Not reported	772 (76.3)	-
Model type		N
Computer vision	860 (85)	-
Signal processing	103 (10.2)	-
Multimodal	18 (1.8)	-
Language	3 (0.3)	-
Other	28 (2.7)	-
Model architecture		Y
Convolutional neural network (CNN)	85 (8.4)	-
U-Net CNN	5 (0.5)	-
Not reported	922 (91.1)	-
Deep learning		Y
Yes	352 (34.8)	-
Not reported	660 (65.2)	-
Evaluation metric		Y
Accuracy	65 (6.4)	Y
Sensitivity	242 (23.9)	Y
Specificity	220 (21.7)	Y
AUROC	110 (10.9)	Y
PPV	66 (6.5)	Y
NPV	54 (5.3)	Y
Other	339 (33.5)	Y

Only 15 devices (1.5%) reported a predetermined change control plan (PCCP), with 73.3% ($n = 11$) of these coming after the introduction of PCCP guidelines in April 2023 (Table 1).

Clinical study variables

Of the 537 devices that reported a clinical study (53.1%), 325 (60.5%) reported a retrospective approach, 75 (14%) a prospective approach, and 11 (2%) reported both. Additionally, 403 of these devices reported a sample size for their clinical study (75%) (Table 1).

Dataset characteristics

In terms of model development, we found that most devices did not report any information on training data source ($n = 944$, 93.3%) nor on the source of their testing data ($n = 764$, 75.5%) (Table 1). Further, only 9.4% of devices ($n = 95$) reported on the training dataset size (number of patients or images), and only 23.2% ($n = 235$) reported on the test dataset size (number of patients or images). We found that only 240 devices (23.7%) reported information on dataset demographics.

Model characteristics

In terms of AI/ML model characteristics, most devices harnessed computer vision ($n = 860$, 85%), with only 3 devices (0.3%) using language as the primary input modality (Table 1). Eighty-five devices (8.4%) harnessed convolutional neural networks, and 352 (34.8%) harnessed deep learning of some sort (Table 1).

Performance metrics

We examined the evaluation metrics reported by each device. We found that most devices did not report performance metrics ($n = 522$, 51.6%). The most frequently reported metrics were sensitivity ($n = 242$, 23.9%) and specificity ($n = 220$, 21.7%), followed by area under the receiver operating

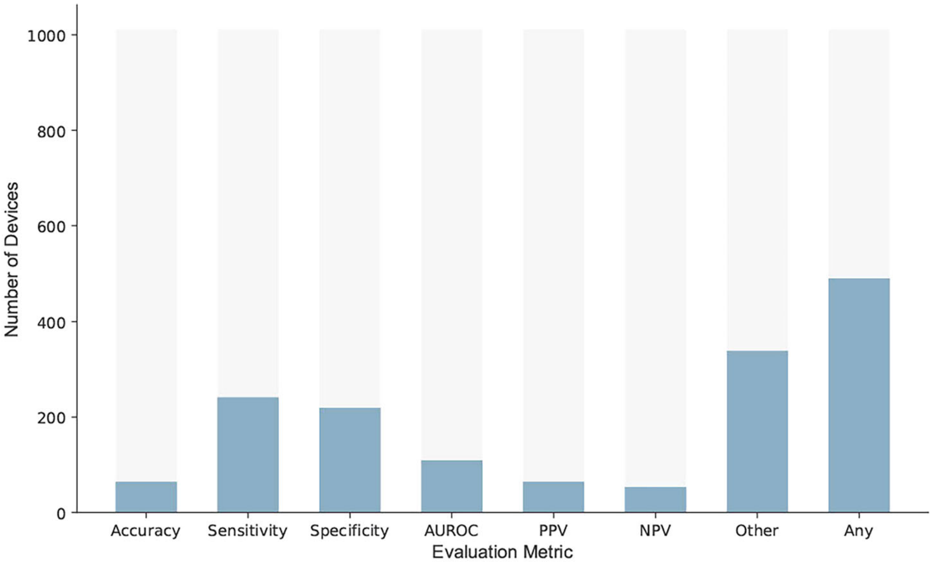

Fig. 1 | Reporting of evaluation metrics among FDA-authorized AI/ML-enabled medical devices. The bar chart shows the number of devices reporting each type of performance evaluation metric at the time of FDA authorization. Blue bars represent the count of devices reporting a given metric, while gray bars represent the total number of devices reviewed. Categories include accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUROC), positive predictive value (PPV), negative predictive value (NPV), and other reported metrics. "Any" indicates reporting of at least one of the listed metrics, including those categorized as "Other.".

Table 2 | Model performance and dataset size metrics for artificial intelligence/machine learning (AI/ML)-enabled devices authorized by the Food and Drug Administration (FDA)

Characteristic	Median (IQR)
Clinical study sample size (no.)	306 (142–650)
Train dataset size (no.)	
Patients	810 (264–5016)
Images	5000 (709–105,000)
Test dataset size (no.)	
Patients	150 (47.5–382)
Images	720 (169–2616)
Evaluation metric (%)	
Accuracy	91.7 (86.4–95.3)
Sensitivity	91.2 (85–94.6)
Specificity	91.4 (86–95)
AUROC	96.1 (89.4–97.4)
PPV	59.9 (34.6–76.1)
NPV	98.9 (96.1–99.3)

characteristic curve (AUROC) ($n = 110$, 10.9%), positive predictive value (PPV) ($n = 66$, 6.5%), accuracy ($n = 65$, 6.4%), and negative predictive value (NPV) ($n = 54$, 5.3%) (Table 1, Fig. 1).

Performance metrics for these devices included median (IQR) sensitivity of 91.2% (85–94.6%), specificity of 91.4% (86–95%), AUROC of 96.1% (89.4–97.4%), PPV of 59.9% (34.6–76.1%), accuracy of 91.7% (86.4– 95.3%), and NPV of 98.9% (96.1–99.3%) (Table 2).

ACTR score

Transparency was quantified using a novel AI Characteristics Transparency Reporting (ACTR) score, a 17-point metric developed in this study (see Methods). The mean ACTR score across all devices was 3.3 out of a total possible 17 points (14 points if a clinical study was not included), with a standard deviation of 3.1. The minimum scoring year was 2008, with a mean of 1.1, and the maximum year was 2023, with a mean of 4 (Fig. 2). The maximum score for any given device was 12, which was found for a single device. The minimum score was 0, which was found for 304 devices (30%).

The linear mixed effects model showed that, on average, scores increased by 0.88 (95% CI: 0.54–1.23) following the publication of the 2021 FDA guidelines, after controlling for whether the device used deep learning and whether the predicate device used AI, and accounting for the correlation in scores within each company (Supplemental Table 1). A more granular analysis of individual metrics revealed that reporting of dataset demographics, and test and training dataset sources and sizes, all differed significantly before and after the release of the 2021 guidelines (all $p < 0.001$) (Supplemental Table 2).

We examined the time between submission and approval for MLMDs cleared via the 510(k) pathway and found that 70.9% of devices ($n = 691$) exceeded the FDA's target review time of 90 days (Fig. 3)[9]. There was a Pearson correlation of 0.15 ($p < 0.001$) between ACTR score and time taken to approve the device.

Discussion

While there is significant national attention on regulating AI/ML technologies, there are few models of regulatory agencies with oversight of these technologies[4,10]. In healthcare, the FDA serves as the primary regulatory authority for AI/ML applications classified as medical devices, making its approach to regulation of these devices an opportunity to inform the national discussion.

As of December 2024, the FDA reported its marketing approval of 1016 medical devices that include an AI/ML technology. Overall, 96.4% of these technologies had been cleared through the 510(k) pathway rather than evaluated as a De Novo submission or through the premarket evaluation pathway (PMA). The 510(k) pathway allows the sponsor to argue that the device under consideration is substantially equivalent to a predicate medical device as a demonstration of its safety and effectiveness, contrasting with the more rigorous De Novo or PMA pathways. Heavy reliance on the 510(k) pathway, which does not inherently require prospective studies, may explain why publicly available clinical evidence is often sparse for AI/ML devices[11,12].

Of the 1016 devices addressed by the FDA, 1012 contained an SSED. On review of these documents, we found substantial gaps in reporting on the performance of the medical device after marketing approval. Overall, reporting on the clinical performance of the underlying medical devices was very limited. Only 53.1% of devices reported a clinical study, and of these, 62.5% were retrospective analyses. Because retrospective studies are more vulnerable to selection bias and dataset leakage (risk of inadvertently including post-outcome data in training) than prospective designs, this evidence base may overestimate real-world effectiveness and limit generalizability[11,12]. The majority, 51.6%, did not report any clinical metrics. Of those that reported metrics, the most frequent were sensitivity (23.9%) and specificity (21.7%), followed by AUROC (10.9%), PPV (6.5%), accuracy

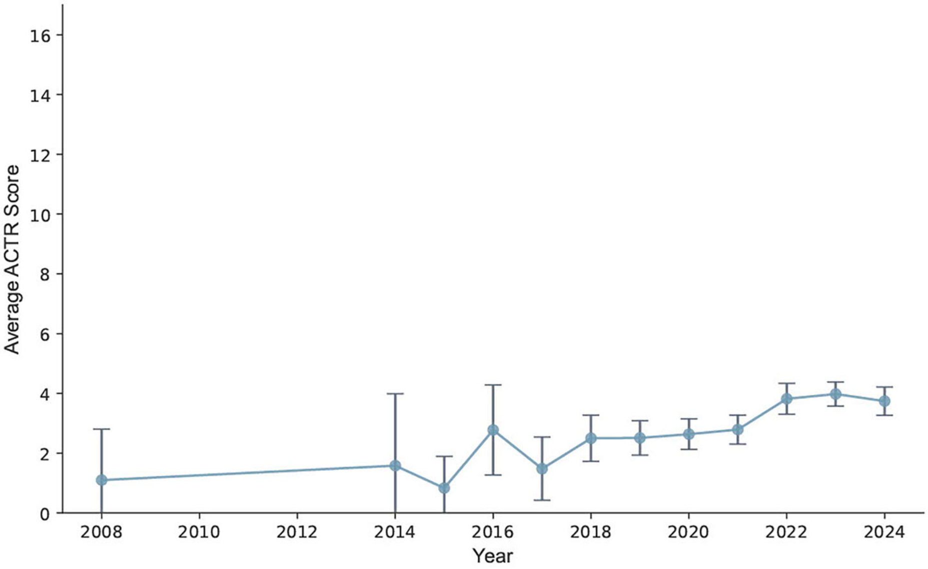

Fig. 2 | Trends in AI characteristics transparency reporting (ACTR) of FDA-authorized AI/ML-enabled medical devices over time. The figure shows the average AI characteristics transparency reporting (ACTR) score for FDA-authorized AI/ML-enabled devices by year of authorization (2008–2024). Only years with at least five devices authorized by the FDA are included. Blue circles represent yearly mean ACTR scores, with blue solid lines connecting the means across years. Error bars indicate 95% confidence intervals, calculated using the standard error of the mean.

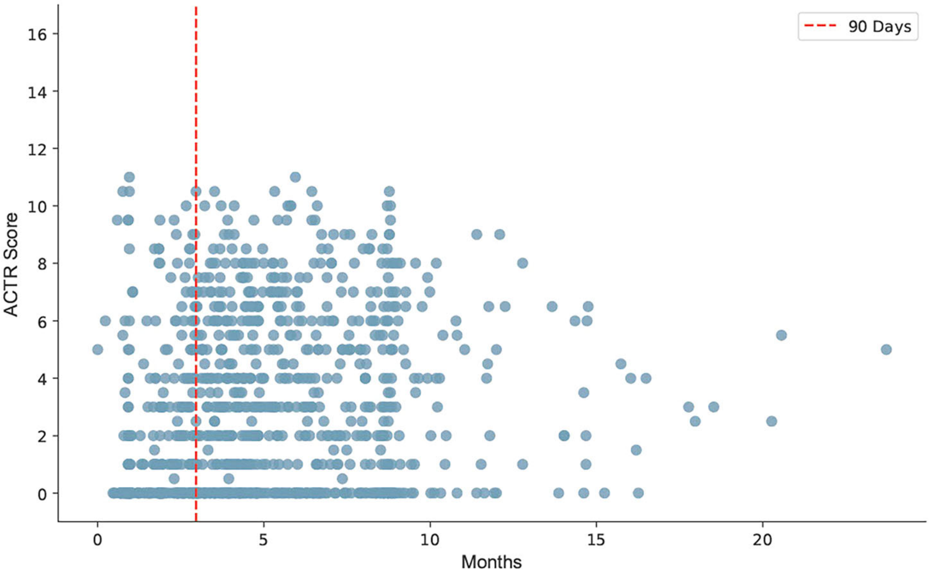

Fig. 3 | Association between AI characteristics transparency reporting (ACTR) score and FDA clearance time for AI/ML-enabled devices. Each blue circle represents a single AI/ML-enabled medical device cleared through the FDA's 510(k) pathway, plotted by its AI characteristics transparency reporting (ACTR) score on the *y*-axis and time to clearance on the *x*-axis. Time to clearance is calculated as the number of months between the FDA's receipt of the submission and the decision date. The red dashed vertical line marks 90 days, the statutory review target for 510(k) submissions.

(6.4%), and NPV (5.3%). The relative infrequency of PPV and NPV reporting is notable because these measures change with pretest probability, so their absence, and the absence of condition prevalence information, limits bedside applicability even when discrimination (e.g., AUROC) appears strong[13].

While the median values associated with these metrics (sensitivities and specificities > 91%, AUROC of 96.1%, and NPV of 98.9%) are impressively high, they should be interpreted with caution because common pre-submission design choices can inflate performance (for example, single-site internal testing, enriched case mixes which can inflate positive predictive value, and post-hoc threshold tuning)[14–16]. Such practices frequently lead to optimism that does not persist on independent, multi-site evaluation; systematic reviews and cross-hospital studies repeatedly find attenuation of performance on external datasets[2,17]. Finally, excellent AUROC or sensitivity does not guarantee an acceptable false-positive burden because predictive values vary with disease prevalence, which was inconsistently reported in device summaries[13,18].

Across all these measures, in addition to the other categories we reported on, we calculated a summary reporting transparency measure, denoted the AI characteristics transparency reporting (ACTR) score. Reporting across all devices was evaluated to an average score of 3.3 out of a total possible 17 points.

The FDA's 2021 guidelines called for transparency on "performance of the model for appropriate subgroups"[7]. While ACTR scores did improve following the release of this documentation, scores remained low relative to the maximum achievable, only increasing by less than a single point (0.88). Scores saw an increase from 2021 to 2022 (2.8 to 3.8) but remained largely stagnant since then (3.7 in 2024) (Fig. 2). The low scores reflect a continued lack of transparency in crucial areas such as performance metrics, training and testing data, and model development. Muralidharan et al., in a review of 692 of these devices, similarly found that only 46.1% provided results of performance studies[19]. Beyond the performance metrics themselves, Wu et al. reported on 130 devices that only 13% of devices considered demographic subgroup performance in their evaluation[20]. Given the growing evidence that medical AI performance can vary across demographic groups—even inferring sensitive attributes from images—under-reporting subgroup analyses increases the risk of inequitable performance in deployment[21,22].

In further analysis, we found that an increased data quality in terms of higher ACTR scores did not correlate with a reduction in time to FDA approval of the device (instead showing a slight positive correlation), suggesting that the current review process does not provide a positive incentive for sponsors to improve the quality of their data in the submission process. Practically, this means there is little time-to-market pressure to enhance transparency, which may help explain why key disclosures remain uncommon despite guidance.

The FDA additionally mandated that "characteristics of the data used to train and test the model" are communicated to stakeholders[7]. However, the SSEDs reviewed here fall short of these guidelines; only 1.8% of devices reported their exact training dataset source, and 3.6% their exact testing dataset source. Additionally, only 4.2% of devices report geographic location of training data sites, and only 16.2% for testing data sites. Dataset provenance is imperative; without it, users cannot assess risks of dataset shift (a shift in the data distribution between training and testing) or selection bias[11]. Similar findings have been reported in reviews of smaller subsets of devices[19,20,23]; Wu et al. found limited reporting of evaluation sites as well as limited geographic diversity in evaluation sites[20]. Additionally, consistent with our findings is that of Ebrahimian et al., who reported in a review of 127 devices that few device summaries report on training or testing dataset sizes or multicenter validation studies[23].

Indeed, lack of transparency in reporting training and testing datasets poses significant challenges to a model's generalizability, raising concerns about domain shift and sampling bias[24–26]. Shick et al. found that without detailed available information on training, testing, and real-world performance, providers expressed hesitancy to use medical AI devices[27]. Additionally, over- or under-representation of certain data in the training set can amplify existing racial and gender disparities, further limiting equitable outcomes[28–31]. Indeed, Warriach et al. report in a recent FDA perspective that AI performance should be monitored in the environment in which it is being used[1], a difficult aim to implement without transparency on how the model was tested and evaluated. Prospective designs and preregistered external, multicenter validation are therefore critical safeguards against optimistic bias[32].

The FDA also instructed device manufacturers to report on "device modifications and updates from real-world performance monitoring"[1]. However, since the PCCP was introduced as a guideline for preemptive reporting of device modifications in April 2023[33], only 3.5% of devices reported a PCCP. This gap is concerning, largely due to the tendency of AI models to experience "model drift," a phenomenon in which AI models are dependent on the data used at the time of training, but quickly degrade in quality as time passes since the last training cycle[3,26,34]. This is especially relevant given that we find an increasing proportion of these devices harnessing deep learning (34.8% all-time, 41% since 2021), making them more prone to drift due to the model sensitivity and complexity[34].

While the U.S. FDA, U.K. MHRA, and Health Canada align on their jointly developed GMLPs and PCCP principles, they diverge in what is legally required prior to approval or clearance.

In the U.K., the first binding element of the MHRA's reform, the new post-market surveillance (PMS) regulations, entered into force on June 16, 2025, now requiring manufacturers to actively track safety and performance, report serious incidents on shorter timelines, and maintain PMS/PSUR (periodic safety update report) reports[35]. In the E.U., the recent AI Act treats most AI devices as "high-risk" systems, requiring strict compliance with mandates on risk management, data governance, technical documentation, transparency, human oversight, and robust post-market monitoring[36,37]. All high-risk devices must undergo third-party conformity assessment by Notified Bodies prior to E.U. market entry[36,37]. Both the E.U. and the U.K. offer stricter frameworks than the U.S., where the FDA's post-market surveillance is limited to adverse event reporting, case-by-case surveillance, and annual reports only for PMA devices. Health Canada further encourages reporting of demographic distribution in training/testing datasets, which was found to be a limitation of the FDA data in this study (only 23.7% of submissions reported dataset demographics)[38].

Empirically, transparency gaps persist in the U.K. and E.U[39,40]. Fehr et al. audited radiology AI products available on the E.U. market and found a median public-transparency score of 29.1% with frequent omissions on training data, ethical safeguards, and deployment caveats[39]. Matthews et al. mapped AI products for digital pathology in the E.U. and U.K., finding sparse and fragmented public evidence, as only 42% had peer-reviewed external validation[40].

In June 2024, the FDA expanded upon 2021 guidance, adding detailed guidelines specifically for "transparency" in MLMDs, encompassing aspects such as device performance, model and dataset characteristics, and underlying technology and machine learning approaches[41]. It is too early to say whether these new guidelines will improve the quality of performance reporting for medical devices.

One path forward for the FDA would be to include a machine-readable "AI Model Card," appended to every publicly available SSED and posted to a public registry. The model card would describe in a standardized fashion the details of the data used for algorithm development, including dataset sources, size, geography, population demographics, disease prevalence, training–test split strategy, evaluation metrics, external validation sites (if any), subgroup results (if any), planned update pathways, and relevant model architecture information[42,43]. Indeed, the FDA's lifecycle draft guidance already includes an example model card, although this approach to structured metadata reporting is not yet mandatory[44,45–47].

With AI/ML, the FDA is being challenged by the development of a novel technology that does not fit well in the current regulatory paradigm for medical devices. Development of these technologies is iterative, so we must understand the performance of the technology when considered for approval/clearance, but also need to monitor the continued performance of the model over time[24,45]. This should include a requirement for adverse event monitoring once the product is marketed. The FDA should consider reporting of suspected adverse events via a pilot National Reporting Indicator (NRI) like that set forth by Wales's Yellow Card NRI, which resulted in an immediate 145% increase in Yellow Card reports[46]. Performance above targets should be incentivized via eligibility for expedited review lanes, while persistent under-performance triggers closer oversight and corrective action. Special attention should be given to concerns over model drift and algorithm updates, requiring site-specific and multi-site validations to ensure generalizability beyond initial premarket studies[2,24].

There are several limitations to our study. First, our review is based on publicly available FDA approval summaries; it is possible that greater transparency is reported in premarket clearance applications to the FDA but not to the public[27]. Second, despite a clear, predefined definition and the resolution of discrepancies via reviewer consensus, the manual review process does still introduce potential subjectivity given the heterogeneity of reporting in the SSEDs. Third, while the ACTR score provides a novel metric for reporting transparency, it does not fully capture the breadth of relevant information or nuances in reporting quality and is not a hierarchical evaluation of the data elements we evaluated.

Despite these limitations, as of December 2024, we have found significant gaps between the FDA guidance on model transparency and the data available to the public on devices approved or cleared for marketing by the FDA. This result suggests that, to date, regulation of AI/ML technologies may have limited utility for the public.

Methods

Data collection

Data were collected from the FDA's publicly available list of "Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices"[8]. This list entails medical devices using AI or ML that have met the FDA's premarket clearance requirements[47]. As of December 20, 2024, there were 1016 devices listed; of these, 1012 included an accessible SSED.

For each device, we manually reviewed the decision summaries to extract the variables of interest (data elements) listed below. At least 1 reviewer manually categorized each device, and at least 1 other reviewer manually verified reporting accuracy. All disagreements were resolved by reaching a reviewer consensus.

To assess device characteristics, we extracted regulatory data, clinical data, categorical data on the AI/ML model use, and performance data on the application.

Regulatory variables

Regulatory data included each device's clearance pathway, submission date, clearance date, associated medical specialty (radiology, cardiovascular medicine, neurology, hematology, gastroenterology/urology, ophthalmology, orthopedic, anesthesiology, general & plastic surgery, physical medicine, obstetrics and gynecology, pathology, toxicology, dental, general hospital, microbiology, or clinical chemistry), predicate device clearance pathway, and predicate device approval date. Each device was either cleared via the 510(k) pathway, authorized for marketing via the De Novo pathway (low-risk class I and II devices without a predicate), or approved via the premarket approval (PMA) pathway (generally class III devices). We also evaluated whether each device reported a predetermined change control plan (PCCP). The FDA defined a PCCP as a document describing "anticipated modifications based on the retraining and model update strategy" that MLMDs could undergo without the necessity for a new premarket review application[33].

Clinical study variables

Each device was assigned to a medical specialty review panel. For those devices that performed clinical studies, we extracted the clinical data collection type (retrospective or prospective) and clinical study sample size.

Dataset characteristics

We extracted information on the datasets used to train, test/and validate the model. These variables included dataset sizes in terms of both the number of patients and the number of images, as well as dataset source reporting as either exact sites, number of sites, region of sites, or not reported. We assessed whether a device reported dataset demographics.

Model characteristics

We also assessed several novel characteristics associated with each device's machine learning model. We classified its model type (computer vision, signal processing, language, multimodal, or other) and model architecture (convolutional neural network (CNN), U-Net CNN, or not reported), as well as whether it used deep learning. Model type was classified based on the input data: images (computer vision), waveforms (signal processing), language (language models), or multiple types (multimodal).

Performance metrics

To evaluate device performance, we extracted several evaluation metrics of interest in ML, including accuracy, sensitivity, specificity, area under the receiver-operating characteristic curve (AUROC), positive predictive value (PPV), and negative predictive value (NPV). We also included a category for "other" evaluation metric for devices reporting on any evaluation metric not included above (such as Dice-Sørensen coefficient, mean squared error, and precision).

ACTR score

To assess comprehensive device reporting on AI/ML model development and performance, we developed the AI Characteristics Transparency Reporting (ACTR) score. The ACTR score is a sum of the reporting elements FDA has identified. A full point is added to the score for reporting each of the following categories: dataset demographics, accuracy, sensitivity, specificity, AUROC, PPV, NPV, other evaluation metric, PCCP, model architecture, training dataset size (patients or images), test dataset size (patients or images), clinical testing, clinical data collection type, and clinical study sample size. Devices can only achieve the latter two points if clinical testing is reported. For training and testing dataset sources, devices received one point for reporting exact sites and half a point for reporting the number of sites or geographic regions.

Data analysis

We calculated descriptive statistics for the percentage of each variable reported or the percentage of each sub-category for categorical variables (Table 1). We also calculated the median and interquartile range for numerical variables related to model performance and dataset size.

We calculated the average ACTR score for each year where a minimum of 5 devices were cleared. To assess the change in reporting transparency (ACTR score) following the 2021 FDA guidelines, we fit a linear mixed effects model with ACTR score as the dependent variable. To account for confounding effects, we included two additional control independent variables: (1) whether a subject device harnessed deep learning, as complex models often necessitate more detailed reporting, and (2) whether a 510(k) device's predicate device also used AI, as prior reliance on an AI-enabled predicate may reduce the need for new extensive reporting. The latter was defined as whether the predicate device also appeared in the FDA's list of cleared MLMDs. We included the company name as a random effect to account for the lack of independence in ACTR scores for devices developed within the same company. We additionally conducted a chi-square test to identify which specific reporting elements contributed most significantly to the change in ACTR score following the 2021 guidelines.

Improved transparency in clinical reporting could be incentivized if it influenced regulatory decisions. We examined the time between submission and clearance for MLMDs cleared by the 510(k) pathway, calculating the Pearson correlation coefficient.

Statistical analyses were performed using Python 3.9 and R 4.4.2.

Data availability

All FDA summaries of safety and effectiveness (SSED) are publicly available and accessible at https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices. All extracted data elements are available from the corresponding author upon reasonable request and approval.

Code availability

All data analysis code is available from the corresponding author upon reasonable request and approval.

Received: 3 July 2025; Accepted: 1 October 2025;
Published online: 17 November 2025

References

1. Warraich H. J., Tazbaz T. & Califf R. M. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. *JAMA* https://doi.org/10.1001/jama.2024.21451 (2024).
2. Youssef, A. et al. External validation of AI models in health should be replaced with recurring local validation. *Nat. Med.* **29**, 2686–2687 (2023).
3. Mashar, M. et al. Artificial intelligence algorithms in health care: is the current Food and Drug Administration regulation sufficient?. *JMIR AI* **2**, e42940 (2023).
4. Gerke, S., Babic, B., Evgeniou, T. & Cohen, I. G. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *npj Digit. Med.* **3**, 1–4 (2020).
5. de Hond, A. A. H. et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digit. Med.* **5**, 1–13 (2022).
6. Joshi, G. et al. FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: an updated landscape. *Electronics* **13**, 498 (2024).
7. Center for Devices and Radiological Health. Good machine learning practice for medical device development: guiding principles. U.S. Food & Drug Administration https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles (2023).

8. Center for Devices and Radiological Health. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. U.S. Food & Drug Administration https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (2024).
9. Center for Devices and Radiological Health. 510(k) submission process. U.S. Food & Drug Administration https://www.fda.gov/medical-devices/premarket-notification-510k/510k-submission-process (2023).
10. Tang, D., Xi, X., Li, Y. & Hu, M. Regulatory approaches towards AI medical devices: a comparative study of the United States, the European Union and China. *Health Policy* **153**, 105260 (2025).
11. Ghassemi, M. et al. A review of challenges and opportunities in machine learning for health. *AMIA Summits Transl. Sci. Proc.* **2020**, 191–200 (2020).
12. Sahiner, B., Chen, W., Samala, R. K. & Petrick, N. Data drift in medical machine learning: implications and potential remedies. *Br. J. Radiol.* **96**, 20220878 (2023).
13. Monaghan, T. F. et al. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. *Medicina* **57**, 503 (2021).
14. Yu, A. C., Mohajer, B. & Eng, J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiol. Artif. Intell.* **4**, e210064 (2022).
15. External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis. *BMC Med. Inform. Decis. Mak.* https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02830-7 (2024).
16. Van Calster, B., Steyerberg, E. W., Wynants, L. & van Smeden, M. There is no such thing as a validated prediction model. *BMC Med.* **21**, 70 (2023).
17. Zech, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
18. Altman, D. G. & Bland, J. M. Diagnostic tests 2: predictive values. *Br. Med. J.* **309**, 102 (1994).
19. Muralidharan, V. et al. A scoping review of reporting gaps in FDA-approved AI medical devices. *npj Digit. Med.* **7**, 1–9 (2024).
20. Wu, E. et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584 (2021).
21. Chen, R. J. et al. Algorithm fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **7**, 719–742 (2023).
22. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D. & Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.* **30**, 2838–2848 (2024).
23. Ebrahimian, S. et al. FDA-regulated AI algorithms: trends, strengths, and gaps of validation studies. *Acad. Radiol.* **29**, 559–566 (2022).
24. Koch, L. M., Baumgartner, C. F. & Berens, P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. *npj Digit. Med.* **7**, 1–11 (2024).
25. Finlayson, S. G. et al. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385**, 283–286 (2021).
26. Kore, A. et al. Empirical data drift detection experiments on real-world medical imaging data. *Nat. Commun.* **15**, 1887 (2024).
27. Shick, A. A. et al. Transparency of artificial intelligence/machine learning-enabled medical devices. *npj Digit. Med.* **7**, 1–4 (2024).
28. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D. & Tzovara, A. Addressing bias in big data and AI for health care: a call for open science. *Patterns* **2**, 100347 (2021).
29. Frasca, M., La Torre, D., Pravettoni, G. & Cutica, I. Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. *Discov. Artif. Intell.* **4**, 15 (2024).
30. Mittermaier, M., Raza, M. M. & Kvedar, J. C. Bias in AI-based models for medical applications: challenges and mitigation strategies. *npj Digit. Med.* **6**, 1–3 (2023).
31. Arora, A. et al. The value of standards for health datasets in artificial intelligence-based applications. *Nat. Med.* **29**, 2929–2938 (2023).
32. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
33. Center for Devices and Radiological Health. Predetermined change control plans for machine learning-enabled medical devices: guiding principles. *U.S. Food & Drug Administration* https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles (2023).
34. Vela, D. et al. Temporal quality degradation in AI models. *Sci. Rep.* **12**, 11654 (2022).
35. GOV.UK. First major overhaul of medical device regulation comes into force across Great Britain. https://www.gov.uk/government/news/first-major-overhaul-of-medical-device-regulation-comes-into-force-across-great-britain (2025).
36. EU Artificial Intelligence Act. Article 72: post-market monitoring by providers and post-market monitoring plan for high-risk AI systems. https://artificialintelligenceact.eu/article/72/ (2025).
37. European Commission. AI Act | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (2025).
38. Health Canada. Pre-market guidance for machine learning-enabled medical devices. https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/application-information/guidance-documents/pre-market-guidance-machine-learning-enabled-medical-devices.html (2023).
39. Fehr, J., Citro, B., Malpani, R., Lippert, C. & Madai, V. I. A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. *Front. Digit. Health* **6**, 1267290 (2024).
40. Matthews, G. A., McGenity, C., Bansal, D. & Treanor, D. Public evidence on AI products for digital pathology. *npj Digit. Med.* **7**, 300 (2024).
41. Center for Devices and Radiological Health. Transparency for machine learning-enabled medical devices: guiding principles. *U.S. Food & Drug Administration* https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles (2024).
42. Schwabe, D., Becker, K., Seyferth, M., Klaß, A. & Schaeffter, T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *npj Digit. Med.* **7**, 203 (2024).
43. Gilbert, S., Adler, R., Holoyad, T. & Weicken, E. Could transparent model cards with layered accessible information drive trust and safety in health AI?. *npj Digit. Med.* **8**, 124 (2025).
44. Center for Devices and Radiological Health. Artificial intelligence-enabled device software functions: lifecycle management and marketing submission recommendations. *U.S. Food & Drug Administration* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing (2025).
45. Matheny, M. E. et al. Enhancing postmarketing surveillance of medical products with large language models. *JAMA Netw. Open* **7**, e2428276 (2024).
46. Deslandes, P. N. et al. Changes in suspected adverse drug reaction reporting via the yellow card scheme in Wales following the introduction of a National Reporting Indicator. *Br. J. Clin. Pharmacol.* **88**, 3829–3836 (2022).
47. Center for Devices and Radiological Health. Device approvals and clearances. *U.S. Food & Drug Administration* https://www.fda.gov/medical-devices/products-and-medical-procedures/device-approvals-and-clearances (2024).

Author contributions

V.M. and K.S. wrote the paper text. V.M., M.N., and K.S. prepared all figures and tables. V.M., A.K., R.B., V.M., M.D.J., and P.S. performed data collection. V.M., B.J., N.S., and K.S. conceptualized the work. All authors reviewed the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-02052-9.

Correspondence and requests for materials should be addressed to Kevin Schulman.

Reprints and permissions information is available at http://www.nature.com/reprints