Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Disclosure of clinical performance evidence in AI/ML device regulation: a cross-sectional audit of 200 US FDA 510(k) device families

Yu Liu

ciciqq1105@hotmail.com

General Hospital of Central Theater Command

Xiang Wang

General Hospital of Central Theater Command

Jun Hou

General Hospital of Central Theater Command

Jianjun Tang

General Hospital of Central Theater Command

Meihua Guo

General Hospital of Central Theater Command

Chaoqun Ma

General Hospital of Central Theater Command

Qingshan Dong

General Hospital of Central Theater Command

Research Article

Keywords: Artificial Intelligence, Device Approval, Documentation, United States Food and Drug Administration, Reproducibility of Results

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11813288/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Regulatory summaries are the main public record of how an AI/ML device was evaluated, yet no protocol with documented inter-rater reliability exists for measuring their disclosure.

Methods

We defined three disclosure attributes with explicit Yes/No/Not-stated/Not-applicable rules; quantitative performance split into a numeric value and a named measure, giving four coded fields. Two coders rated 258 US families and 48 China summaries separately; 200 US families passed the identity screen, and the China module came from 616 summaries. We applied the protocol to 200 US FDA families sampled within review panels, with design-weighted estimates, and report China descriptively.

Results

Reliability was high for identity (0.916), quantitative-performance (0.932), independence (0.955) and named-measure coding (0.888); attribution was lower (0.742). Of 200 US families, 158 (79.6% weighted) named a quantitative measure, 156 (80.1%) attributed evidence to a tested version, 115 (58.5%) reported a numeric estimate, and 43 of those 115 (40.3%) documented independent evaluation. In a secondary composite, 38 met all three criteria (21.3%); the split between explicit failure (96) and insufficient information (66) shifted to 37 and 125 under an alternative coding of non-numeric statements.

Conclusions

Of 115 families reporting a numeric estimate, 43 (40.3% design-weighted) documented independent evaluation. A requirement to report the three domains we coded, as four categorical fields, would let a reader check what a public summary reports without prescribing a performance threshold.

Introduction

Medical devices across specialties incorporate artificial intelligence and machine learning (AI/ML) algorithms [1, 2]. Approval patterns have changed over time [3, 4], and reviews have examined Chinese applications [5]. Unlike conventional devices, AI/ML components derive decision rules from training data and can vary across deployment settings [6], so transparent reporting matters to clinicians, health systems and researchers [7].

US and Chinese regulators require manufacturers to submit performance data for premarket clearance or approval [8], and both publish summaries: the FDA through 510(k) summaries, China’s NMPA through

CMDE review reports. Previous work documented limitations in device evaluation [6] and in AI-versus-clinician study design [7], validation gaps linked to recalls [9], and deployment bias [10]. Studies have examined annotation and ground-truth disclosure [11], the observability of authorised devices [12], and how prevalence changes the proportion of positive results that are false [13]. Few used reproducible coding criteria to audit which evidence attributes summaries disclose.

We developed a coding protocol with prespecified definitions and tested its reliability before application. We audited three attributes in US FDA 510(k) summaries: independent evaluation, quantitative performance metrics, and component-level attribution. In Chinese summaries we assessed the two comparable attributes; attribution was assessed only in the US cohort. For US families only we applied a composite of the three criteria and report it as a secondary analysis. China's structural analogue of attribution duplicated b2, so it was not treated as an independent criterion and no China composite was derived; that module is descriptive. For the three checks we audited, the domains map to four coded fields: a numeric value, documented independence conditional on that value, a named measure, and the tested version. A fuller template would add interval and denominator, reference standard and population; only interval and sample-size reporting were audited separately, in China.

We asked three questions. First, can coders reproducibly measure three disclosure attributes in regulatory summaries using a prespecified protocol? Second, how often does each attribute appear in US FDA 510(k) summaries? Third, how much do composite findings depend on the rule for combining attributes?

Materials and Methods

Study design and regulatory context

We conducted a cross-sectional audit of publicly available US FDA 510(k) summaries and Chinese NMPA approval summaries. Independent coders applied predefined verifiability criteria. AI reporting guidelines and evaluation frameworks [14, 15, 16] and reviews of regulatory evidence requirements [17] informed the audit framework. The device family was the unit of analysis. We define the three attributes below.

We registered the study on the Open Science Framework (https://osf.io/etwqr) before coding; the dated amendment record is archived with the registered materials (osf.io/ck8yf). Retrieval closed on 21 October 2026, and an eligibility check before coding reduced the frame from 601 to 505 families. We redrew the 258-family sample from the corrected frame with the registered design and seed. Coding began on 23 October 2026; 200 families passed the identity screen.

US cohort: sampling and eligibility

The sampling frame was drawn from the FDA 510(k) public database and comprised families with a decision date between 2 January 2020 and 21 October 2026. A family comprised all clearances sharing an applicant and device name, so one family could span several summaries; coding was at family level. Removing 96 families whose clearances all preceded 2 January 2020 left 505 eligible families, from which we sampled 258 for dual coding (Fig. 1). Of the 258 sampled families, 58 did not pass the identity screen (non-AI/ML or unclassified) and were excluded; all estimates apply to the 200 included families.

China cohort: enumeration and eligibility

We enumerated all publicly available approval summaries on the National Medical Products Administration (NMPA) Center for Medical Device Evaluation (CMDE) portal for devices approved between 2022 and 2025 (n = 616). Two independent coders screened every report. The final China cohort comprised 48 AI/ML devices.

Two coders classified 54 of 616 reports as AI/ML, 341 as non-AI and 221 as not stated; adjudication excluded 39 not-stated reports with no attributable AI component, leaving 182. Full-text screening of these identified 14 candidates but added no devices. We rechecked seven AI/ML reports describing in vitro reagents or non-learning software and excluded six, leaving 48.

Verifiability constructs and operationalisation

Verifiability refers to three disclosure attributes, not to whether a device can be verified. Quantitative performance (b1_quant) was Yes if the summary reported at least one numeric performance value for the claimed output, No for attainment statements, qualitative descriptions or predicate parameters alone, and Not Stated (NS) if performance evaluation was not mentioned. Independence (b1_indep) was Yes if training, tuning and feature selection did not use the evaluation data, No for explicit non-independence, and NS for design descriptions that did not relate the data to development; when b1_quant was No or NS, b1_indep was Not Applicable. Quantitative metrics (b2) were Yes for a named measure, No if explicitly unavailable, NS otherwise. Component attribution (b3) was Yes if evidence could be linked to a single evaluation of the AI component and a specific version, reflecting predetermined change control documentation [18]; b3 applied to the US cohort only. Coders assigned S2 if any of the four fields was explicitly No, S1 if all four were Yes and traceable to one evaluation, and S3a otherwise. Any explicit No took precedence over NS or NA. Supplementary Table S3 gives the full coding rules and denominators.

For the China cohort, we used the cleaned full text to record whether each summary reported a confidence interval or equivalent interval estimate and whether it reported a countable sample size. The same two coders independently rated both items as Y, N or unclear without seeing each other’s ratings or the automated extraction results. We report China estimates separately as unweighted counts across

disclosed summaries. We excluded five calibration summaries from agreement statistics, leaving 43; a third reviewer adjudicated one discrepant cell.

Statistical analysis and agreement

For the US cohort we used the registered stratified ratio estimator with FDA review panels crossed with clearance year as strata; each weight was the number of eligible families in a stratum divided by the number sampled in that stratum. We calculated design-based 95% confidence intervals (derivation in the archived analysis files, osf.io/ck8yf) and estimated reliability in the 258-family sample, of which 200 contributed outcome estimates.

Families were classified by the composite rule signed on 29 October 2026, before the 200-family recode: any explicit No in the four fields counted as S2. S1 was later tightened to require all three domains traceable to one evaluation; this dated requirement applies to S1 only. A post hoc sensitivity analysis treated b1_quant = No as insufficient information. A count-based transition table is in the supplement.

Supplementary Table S3 gives the operational definitions and coding categories; the composition of the excluded sets is recorded in the archived dataset (osf.io/ck8yf).

We followed the STROBE statement for cross-sectional studies .

Results

Of the 200 US AI/ML device families, 115 (design-weighted 58.5%, 95% CI 50.9–66.1) reported at least one numeric performance estimate for the claimed output. Of these, 43 (43/115; design-weighted 40.3%, 95% CI 29.5–51.1) documented that evaluation data were independent of development. Across all 200 families, 158 named a quantitative measure and 156 attributed evidence to a tested component version. Overall, 38 families met all three disclosure criteria (19.0% crude; design-weighted 21.3%, 95% CI 14.1–28.4). Of the remainder, 96 did not meet at least one prespecified reporting criterion (48.0% crude; 46.7% design-weighted), and 66 provided insufficient information (33.0% crude; 32.1% design-weighted). In the post hoc sensitivity analysis, 37 did not meet at least one criterion (18.5% crude; 17.4% design-weighted) and 125 provided insufficient information (62.5% crude; 61.3% design-weighted). Supplementary Table S3 gives definitions and denominators for b1, b2, b3 and S1/S2/S3a.

Protocol development and reliability

We developed the protocol in three calibration stages before the main analysis. Before coding, we defined the attributes and specified that independence was not applicable without a reported numeric estimate. After reviewing the coded distributions, we amended the precedence rule and report both versions. Two coders independently rated every family and adjudicated each discrepancy. We assessed agreement across the 258 sampled families using Cohen’s kappa and Gwet’s AC1 [19]; the two split b1

fields were evaluated in the 200-family set. Results were: identity classification 0.916 (AC1 0.962); quantitative performance 0.932 (0.953); independence 0.955 (0.961); named measure 0.888 (0.922); attribution 0.742 (0.823). Attribution was lower in the 195-family consensus subset (0.587; AC1 0.834), indicating that the component-version item was the least reproducible element. Agreement details for both cohorts, including the China b1 distribution, are in Supplementary Table S4.

Protocol amendment and its effect

The signed rule treated non-numeric performance statements as an explicit No; summaries silent on performance evaluation received Not Stated, and four families were coded that way. During coding, we found that this rule conflated summaries stating that independence had not been established with those reporting no number. The primary analysis retains the signed rule. A post hoc sensitivity analysis classifies the latter group as having insufficient information. After coding began, we tightened S1 to require that readers could trace all three domains to a single evaluation. This change, recorded in a dated amendment, did not alter the S1 count in these data: 38 families met all four criteria before and after tightening. We documented both changes so readers can identify which decisions preceded the results.

The S2 rule was signed on 29 October 2026 at 10:26, before the 200-family recode was run: any explicit No in b1_quant, b1_indep, b2 or b3 placed a family in S2. The recode was completed after that signature and produced 38 S1, 96 S2 and 66 S3a families. A later dated amendment tightened S1 to require one evaluation traceable across all three domains; it left the S1 count at 38 and applies to S1 only. A post-hoc sensitivity analysis, in which b1_quant=No is treated as insufficient information rather than as an explicit No, moved 59 families from S2 to S3a and gave 38 S1, 37 S2 and 125 S3a. The signed rule gives the reported classification, with the alternative coding presented as a sensitivity analysis.

Cohort characteristics

The US cohort comprised 200 AI/ML device families cleared through the 510(k) pathway between 2020 and 2026. These came from a stratified random sample of 258 families drawn from a frame of 505; 200 families met the identity criteria, and we report design-weighted estimates alongside crude proportions. Agreement on the three-category identity classification across the 258 families was κ=0.9158 (95% CI 0.852–0.973; 250 concordant). We arbitrated the 8 discrepant cases: Five were AI/ML, 2 non-AI/ML, and 1 unclassifiable. After arbitration, 200 families met the AI/ML inclusion criteria (195 concordant + 5 arbitrated) and formed the main analysis set. (Figure 1, Table 1)

From 616 screened summaries, we identified 48 NMPA-approved AI/ML devices for which CMDE published review reports between 2022 and 2025. These formed the China cohort. We reclassified seven devices as in vitro diagnostic reagents, excluded six, and retained one (CSZ2300184).

Independent evaluation (b1)

In the US cohort, 115/200 summaries reported at least one numeric performance value for the claimed output (b1_quant=Yes; design-weighted 58.5%, 95% CI 50.9–66.1). Another 81 reported only attainment statements, qualitative descriptions, or predicate parameters (b1_quant=No; 39.6%, 32.1–47.1). Four provided no evaluable performance statement (b1_quant=NS; 1.9%). These percentages are design-weighted; crude proportions were 57.5%, 40.5%, and 2.0%, respectively. (Table 2)

Of the 115 families reporting a numeric performance value, 43 stated that development did not use the evaluation data (b1_indep=Yes; weighted 40.3%, 95% CI 29.5–51.1). Another seven explicitly reported non-independence (5.6%), and 65 did not describe the relationship (54.1%). Across the full sample, the corresponding estimates were 23.6%, 3.3%, and 31.7%; independence was not applicable to 41.5%.

In the China cohort, 38/48 summaries reported at least one numeric performance value (b1_quant=Yes; 79.2%, 95% CI 65.7–88.3), and 10 reported none (20.8%, 11.7–34.3). Of the 38, none stated that the evaluation data were independent of development, and 37 described design features without relating the data to development. The 38th concerned a predecessor device. For the ten reports without numeric values, b1_indep was not applicable.

Quantitative metrics (b2)

In the US cohort, 158/200 summaries reported at least one quantitative performance metric (b2=Yes; 79.6% weighted, 95% CI 75.4–83.8). Another 19/200 explicitly stated that none was present (9.9%, 6.7–13.0), and 23/200 did not address metrics (10.5%, 7.4–13.7).

In the China cohort, 39/48 summaries (81.3%, 95% CI 68.1–89.8) reported quantitative metrics (b2=Yes). None (0/48) explicitly stated that metrics were absent (b2=No). Nine (9/48, 18.8%) did not address metrics (b2=NS). Raters agreed on b2 in 45/48 cases (κ=0.778 and AC1=0.927). For the unsplit b1 item, both coders classified 45 of 48 summaries as not stated; agreement was 45/48 with AC1 0.934, and kappa (0.377) is not interpretable because the marginals were near-constant (Supplementary Table S4).

Of the 9 China cohort summaries coded b2=NS, 3/9 (33.3%) reported sample size or data provenance but no performance values. The remaining 6/9 reported neither sample size nor performance values.

Component attribution (b3, US cohort) and structural consistency (b3a, China cohort)

In the US cohort, 156/200 summaries provided unambiguous component-level attribution (b3=Yes; design-weighted 80.1%, 95% CI 76.0–84.2). Another 11/200 explicitly reported no such attribution (b3=No; 4.1%, 2.3–5.9), and 33/200 did not address it (b3=NS; 15.8%, 11.9–19.6).

In the China cohort, b3a recorded only whether numeric values appeared in the performance paragraph. Positive cases matched b2=Yes exactly (39/48), so b3a measured document structure rather than a

separate aspect of verifiability. Agreement was 46/48 (κ=0.850, AC1=0.942). We used b3a to confirm coding consistency and did not interpret its prevalence separately.

Composite classification (secondary analysis)

We defined the composite classification only for the US cohort and report it as secondary; the attribute estimates do not depend on it. Under the signed rule, 38 families met all three criteria (design-weighted 21.3%, 95% CI 14.1–28.4); 96 (48.0% crude; 46.7% weighted) failed at least one criterion and 66 (33.0%; 32.1%) provided insufficient information. Of the 43 b1_indep=Yes families, 38 were S1, 3 S2 and 2 S3a. A post hoc sensitivity analysis treated b1_quant=No as insufficient information and moved 59 families, giving 38 S1 (21.3%), 37 S2 (18.5% crude; 17.4% weighted) and 125 S3a (62.5% crude; 61.3% weighted) (Table 3).

In the sensitivity analysis, among the 125 S3a summaries, 123 of 125 (98.4%, crude) did not document that evaluation data were independent of development. Sixty-three reported no numeric performance estimate, so independence was not applicable. Sixty reported a numeric estimate without relating the evaluation data to development. The remaining 2 documented independence but did not name the measure (b2=NS).

China cohort: sample size and data provenance reporting

We assessed parameter completeness separately in all 48 China summaries. A confidence interval appeared in 38 (79.2%, 95% CI 65.7–88.3) and a countable sample size in 42 (87.5%, 75.3–94.1). Five (10.4%) described a multicenter study without specifying the number of participants, cases or images. Six reported neither a confidence interval nor a countable sample size. Of the 38 reporting a confidence interval, 37 (97.4%, 95% CI 86.5–99.5) described design features without stating whether evaluation data were independent of development; the remaining summary concerned a predecessor device. None of the 38 documented independence (Supplementary Table S2).

Variation by review panel

Disclosure varied by review panel (Table 4). Radiology families were more likely than all other panels combined to report a numeric estimate (80/124, 64.5% crude; 66.0% weighted, vs 35/76, 46.1% crude; 45.5% weighted; Fisher p=0.012, Holm-adjusted p=0.024) and to meet all three criteria (36/124, 29.0% crude; 31.9% weighted, vs 2/76, 2.6% crude; 2.6% weighted; p<0.001, Holm-adjusted p<0.001). Gastroenterology-Urology had the highest crude proportion reporting a numeric estimate (9/12, 75.0%); no family in Neurology or Gastroenterology-Urology met all three criteria. These subgroup comparisons were not prespecified and are exploratory; Holm adjustment was applied across the two comparisons within each outcome and is reported alongside the unadjusted values.

Discussion

Public regulatory summaries documented the three attributes unevenly: 158 families named a quantitative measure, 156 attributed evidence to a tested version, 115 reported a numeric estimate, and 43 of those documented independent evaluation. These constructs capture what a summary discloses about independence, metrics and attribution; they do not assess evaluation adequacy or clinical effectiveness.

In the US cohort, 43 of 115 documented independent evaluation; 65 did not describe the relationship and seven reported non-independence. The attribute estimates do not depend on the composite rule, which was secondary: classifying statements without a numeric value as failures or as insufficient information moved 59 families. Incomplete disclosure is consistent with absent requirements or inconsistent implementation; an S3a or NS code means the summary did not address the question, not that evaluation was absent. The panel difference locates where disclosure is weakest and could inform panel-specific guidance rather than a uniform requirement.

One NMPA summary states, in translation, that sensitivity and specificity of software-assisted reading met the expected target, without values, sample size, reference standard, or any statement on prior use of the data for training. This gap concerns the public record, not necessarily the study.

Previous analyses identified incomplete reporting in surgical AI studies [20] and trials of large language models [21], as well as limitations in the evidence used to evaluate medical AI devices [6, 7]. This audit uses reproducible coding rules and two raters to produce jurisdiction-specific estimates. S2 marks a family assigned No on at least one field under the signed rule; S3a marks a family whose summary did not address a criterion. Neither label describes the underlying evaluation. Implementation-readiness frameworks for clinicians address a related but distinct question [22].

We prespecified the China module as secondary and descriptive. Both cohorts used comparable fields for b1_quant and b2, but no China summary with a numeric estimate documented independence (0/38), so that comparison is one-sided. b3 applies only to the US cohort; in China, b3a reflects document structure rather than a separate aspect of verifiability. The module generates hypotheses but does not support comparisons across jurisdictions.

Under 21 CFR 807.92 [23], US reporting requirements depend on whether substantial equivalence rests on performance data. If it does, the 510(k) summary must describe the nonclinical and clinical tests supporting the determination, the subjects tested, the data obtained and the conclusions drawn. The provision does not require a statement on whether evaluation data were independent of development, an interval estimate, or identification of the tested component version. Because the submitter writes the summary, an omission marks a gap in the document the submitter must supply.

The Chinese registration route requires a different interpretation. Where applicants rely on a clinical trial, they must submit the trial protocol, ethics approval, informed consent and trial report [8]. The reviewing

authority publishes its own technical review report, for which we found no published specification. All 48 reports described the evaluation route and validation materials, and most reported the number of patients or centers.

Regulators could require applicants to complete standard tables on evaluation independence, quantitative metrics and, where applicable, component version identifiers, consistent with existing AI reporting guidelines [14]. For the three checks we audited, four coded fields were sufficient: a numeric performance value, documented independence conditional on that value, a named measure, and the tested component version. A fuller template would add the interval and denominator, the reference standard and the population evaluated. Our data describe current practice but do not show that templates would improve outcomes. Public summaries alone cannot support evidence synthesis; reviewers should seek peer-reviewed validation studies, post-market evidence or manufacturer reports when available [6, 7]. Procurement and health technology assessment teams can use the three domains to guide requests. A missing item warrants a question, not a conclusion that the device failed.

Disclosure does not establish evaluation quality, but it determines whether readers can independently assess it. Regulators could require three attributes without setting a performance threshold: the evaluation dataset identifier and its separation from development, a value with its interval and denominator, and the tested component version. Disclosure requires documentation already generated during evaluation, not new studies. Without this requirement, the documents linking marketing authorization to clinical use do not permit verification; readers cannot distinguish a device that lacked independent evaluation from one whose summary omitted it.

When neither summary states how evaluation data related to development, a clinician cannot compare two cleared devices on that basis, and a procurement committee lacks a common field. If the summary omits the tested version, readers cannot link an update to the evidence supporting clearance.

Limitations

First, the two jurisdictions used different sampling strategies and periods; we describe them separately and treat China as secondary and descriptive. The weighted estimates apply only to the 505-family US frame (decisions from 2 January 2020 to 21 October 2026), not to all 510(k) clearances.

Second, component attribution (b3) and the composite were calculated only for the US cohort; in China, b3a duplicated b2 and could not serve as a third criterion.

Third, outcome coding began on 23 October 2026. We registered the attribute definitions before coding; the precedence rule was signed on 29 October 2026, after initial coding had begun and before the 200-family recode; one criterion was tightened later. A post hoc sensitivity analysis is reported separately.

Fourth, we audited public summaries only. Neither NS nor S3a distinguishes absent evidence from withheld evidence, so our estimates describe disclosure, not the underlying evidence.

Conclusion

In this audit of 200 US 510(k) summaries for AI/ML device families, disclosure was uneven at the attribute level: most summaries named a quantitative performance measure (158, 79.6% design-weighted) and attributed evidence to a tested component version (156, 80.1%), 115 (58.5%) reported a numeric estimate, and only 43 of those 115 (40.3%) stated that evaluation data were independent of development. The composite classification was secondary, and its boundary between explicit failure and insufficient information depended on how statements without a numeric value were coded. Because these attributes are documentary, regulators could require the three domains, operationalised as four categorical fields, without setting a performance threshold. Disclosure determines whether a reader can check a claim, not whether the device performs as claimed. Both agencies could adopt these four fields without setting a performance threshold or altering the evidence standard for authorization.

Declarations

Ethics approval: Not required; the study analysed public regulatory documents and involved no human participants. Use of generative AI: tools were used for literature search, language editing, translation, drafting analysis scripts and independent methodological review; the authors evaluated all suggestions, made all design and analysis decisions, and verified all counts and citations. Figure 1 was produced by the authors from the audited data; no third-party image or material requiring permission is reproduced.

Acknowledgments

The authors received no funding, and no one outside the author list contributed to this work.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

Conflict of interest: The authors declare no financial or non-financial competing interests.

Author Contributions

Yu Liu conceptualized and designed the study, performed formal analysis, and drafted the manuscript. Xiang Wang and Jun Hou contributed to investigation and validation; Yu Liu and Xiang Wang independently coded all documents, and Jun Hou adjudicated disagreements. Jianjun Tang and Meihua Guo curated data and wrote analysis scripts; Chaoqun Ma produced the figure; Qingshan Dong supervised. All authors approved the final version and accept responsibility for its content.

Data Availability

The audited documents are public 510(k) summaries from the US Food and Drug Administration and publicly disclosed NMPA documents, both retrievable from the respective agency websites. The coded dataset, codebook, analysis scripts and amendment record are deposited in the Open Science Framework (OSF) repository at https://osf.io/ck8yf. The study protocol was registered on OSF at https://osf.io/etwqr.

References

1. Mohan Babu A, Jeshurun Nehemiah A, Jagadeep V, et al. Deep learning and generative AI for medical imaging and clinical decision support systems: a structured critical review. Front Digit Health. 2026;8. 10.3389/fdgth.2026.1890072.
2. Hussain A, Guni A, Gandhewar R, et al. Is the FDA regulation of cardiology AI devices supporting cardiovascular innovation: a scoping review. Heart. 2025;112(6):300–6. 10.1136/heartjnl-2025-326307.
3. Liu Y, Yu W, Dillon T. Regulatory responses and approval status of artificial intelligence medical devices with a focus on China. npj Digit Med. 2024;7:255. 10.1038/s41746-024-01254-x.
4. Golshani P, Joseph MS. Artificial intelligence and machine learning-enabled medical devices: persistent specialty concentration and the care-delivery gap (1995–2025). Cureus. 2026. 10.7759/cureus.112583.
5. Wang Z. Artificial intelligence in Chinese healthcare: a review of applications and future prospects. Biomed Eng Lett. 2025;15(6):1065–72. 10.1007/s13534-025-00515-2.
6. Wu E, Wu K, Daneshjou R, et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021;27:582–4. 10.1038/s41591-021-01312-x.
7. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689. 10.1136/bmj.m689.
8. National Medical Products Administration. Measures for the Administration of Medical Device Registration and Filing (Order No. 2021;47). Beijing: NMPA.
9. Lee B, Kramer P, Sandri S, et al. Early Recalls and Clinical Validation Gaps in Artificial Intelligence–Enabled Medical Devices. JAMA Health Forum. 2025;6(8):e253172. 10.1001/jamahealthforum.2025.3172.
10. Kalyanpur A, Mathur N. Clinical Validation, Bias, and Ethical Deployment of Artificial Intelligence in Imaging. Indian J Radiol Imaging. 2026;36(04):470–9. 10.1055/s-0046-1825797.
11. Youssef A, Fronk D, Grimes JN et al. Beyond the Black Box: Avenues for Transparency in Regulating Radiological AI/ML-enabled SaMD via the FDA 510(k) Pathway. medRxiv [Preprint]. 2024. 10.1101/2024.07.12.24309602

12. Lavinda O. Premarket transparency and postmarket observability in FDA-authorized AI-enabled medical devices: a source-linked cross-sectional study. Res Square [Preprint]. 2026. 10.21203/rs.3.rs-10475650/v1.
13. Sparnon E, Stevens K, Song EC et al. The false positive paradox: Examining real-world clinical predictive performance of FDA-authorized AI devices for radiology using clinical prevalence. medRxiv [Preprint]. 2026. 10.64898/2026.03.25.26349197
14. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904. 10.1136/bmj-2022-070904.
15. Mora-Delgado J, Rivas Jiménez Vc, Lojo-Cruz C. The PALLI-AI Framework: An Evidence-Informed Checklist for Designing and Reporting Trustworthy Artificial Intelligence Tools in Palliative and End-of-Life Care. Healthcare. 2026;14(18):2998. 10.3390/healthcare14182998.
16. Khoruzhaya AN, Varyukhina MD, Erizhokov RA, et al. MEDAI-LLM-SUMM: a reporting checklist for medical text summarization studies using large language models. Front Digit Health. 2026;8. 10.3389/fdgth.2026.1761601.
17. Saini M, KC G, Williams AJ, et al. Therapeutic Innov Regul Sci. 2025;60(2):393–422. 10.1007/s43441-025-00896-7. Regulatory Challenges and Opportunities: A Review of U.S. Food and Drug Administration-Approved Artificial Intelligence and Machine Learning-Enabled Cardiovascular Devices.
18. Dayma K, Patel P, Hildreth K et al. Predetermined Change Control Plan Adoption and Documentation Transparency in U.S. Food and Drug Administration–cleared Radiology Artificial Intelligence/Machine Learning Devices. Radiology: Artificial Intelligence. 2026;8(5). 10.1148/ryai.260385
19. Hallgren KA. Tutorials Quant Methods Psychol. 2012;8(1):23–34. 10.20982/tqmp.08.1.p023. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial.
20. Gandhewar R, Guni A, Tandon D, et al. Surgical Artificial Intelligence Innovation Trends and Regulatory Insights. Ann Surg. 2026. 10.1097/sla.0000000000007229.
21. Chen D, Blayne A, Ning T, et al. Reporting completeness of clinical trials evaluating large language model interventions in medicine: systematic review. BMJ. 2026;395:e100855. 10.1136/bmj-2026-100855.
22. Nafie S, TRIAGE-AI. A clinician-facing implementation-readiness framework for evaluating artificial intelligence in cancer pathways. Int J Med Informatics. 2026;220:106623. 10.1016/j.ijmedinf.2026.106623.
23. United States Food and Drug Administration. 21 CFR 807.92: Content and format of a 510(k) summary. Code of Federal Regulations, Title 21, Part 807, Subpart E. Washington, DC: Office of the Federal Register.

Tables

Table 1. Characteristics of the US and China cohorts

Characteristic	US cohort	China cohort
Sampling frame	505 device families (decision dates 2020-01-02 to 2026-10-21)	616 screened summaries (approvals 2022–2025)
Sampling method	Stratified sample of the 505-family frame: random sampling within large strata and full enumeration of small strata; estimates were design-weighted with the registered stratified ratio estimator (see Methods)	Full enumeration of publicly disclosed AI/ML approvals; no sampling, so estimates were unweighted
Initial screening	258 device families dual-coded for AI/ML identity	54 candidate AI/ML devices identified
Identity agreement	κ=0.9158 (95% CI 0.852–0.973) for three-category classification across 258 families; κ=0.8589 (0.809–0.909) for five-level classification	κ=0.8886 (95% CI 0.8548–0.9225) across 616 reports (crude agreement 93.67%)
Exclusions	49 non-AI/ML (47 concordant, 2 via arbitration) and 9 unclassifiable (8 concordant, 1 via arbitration)	7 in vitro diagnostic reagents (6 excluded; 1 retained: CSZ2300184)
Main analysis set	200 AI/ML device families	48 AI/ML devices

Table 2. Verifiability domain prevalence: US cohort (n=200)

Domain	Yes, n (weighted %)	No, n (weighted %)	Not stated, n (weighted %)
b1_quant: Numeric performance reported	115 (58.5)	81 (39.6)	4 (1.9)
b1_indep: Evaluation data independent†	43 (40.3)	7 (5.6)	65 (54.1)
b2: Quantitative metrics	158 (79.6)	19 (9.9)	23 (10.5)
b3: Component attribution	156 (80.1)	11 (4.1)	33 (15.8)

Design-weighted estimates from the registered stratified ratio estimator; the unweighted denominator is 200 families. †Percentages for evaluation-data independence are conditional on a reported numeric estimate (denominator 115); 85 families were coded not applicable.

Table 3. Verifiability classification: US cohort only (n=200)

Category	Definition	n (weighted %)
S1	All criteria met (b1_quant=Yes, b1_indep=Yes, b2=Yes, b3=Yes; single evaluation source)	38 (21.3)
S2	At least one criterion explicitly not met (including b1_quant=No)	96 (46.7)
S3a	Insufficient information: no explicit failure	66 (32.1)

*Design-weighted; unweighted denominator 200 families. Signed rule: S1 38 (21.3%), S2 96 (46.7%), S3a 66 (32.1%). Sensitivity: 38/37/125.

Table 4. Disclosure by FDA review panel (US cohort, n=200)

Panel	n	Numeric estimate, n (crude %; weighted %)	Met all three criteria, n (%)
Radiology	124	80 (64.5; 66.0)	36 (29.0)
Cardiovascular	16	9 (56.2; 56.2)	1 (6.2)
Neurology	14	5 (35.7; 35.7)	0 (0.0)
Gastroenterology-Urology	12	9 (75.0; 75.0)	0 (0.0)
Other panels	34	12 (35.3; 34.6)	1 (2.9)

Radiology versus all other panels: numeric estimate 64.5% vs 46.1% (Fisher p=0.012; Holm-adjusted p=0.024); met all three criteria 29.0% vs 2.6% (p<0.001; Holm-adjusted p<0.001). Crude and design-weighted percentages are shown where they differ; sampling strata were review panel crossed with clearance year, so a within-panel weighted estimate need not equal the crude proportion. Panel subgroups were not prespecified and are exploratory.

Figures

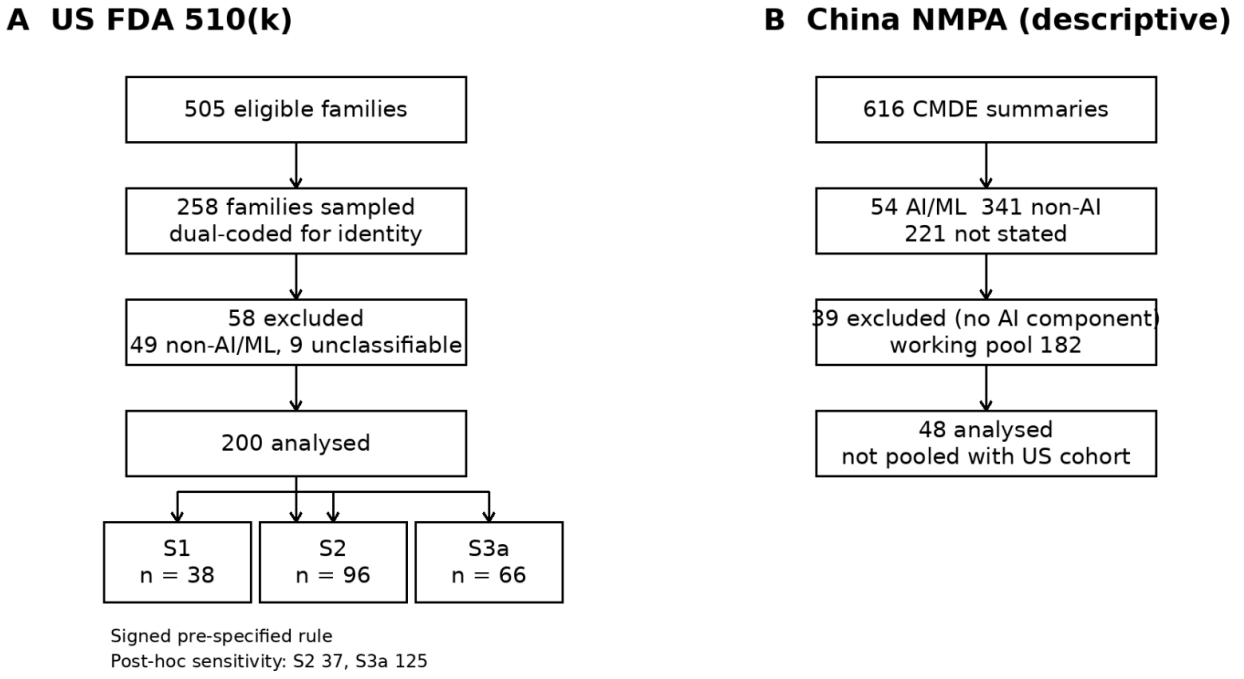

Figure 1

Study flow diagram. The panels show the separate US FDA and China analyses. Of 505 eligible US device families, 258 formed a stratified random sample, of which 58 were excluded at the identity screen (non-AI/ML or unclassified) for dual AI/ML identity coding. The analysis excluded 58 families: 49 non-AI/ML (47 concordant, 2 by arbitration) and 9 unclassifiable (8 concordant, 1 by arbitration). The signed rule divided the 200-family main analysis set into S1 (n=38), S2 (n=96), and S3a (n=66). In China, dual identity coding classified 616 CMDE summaries: 54 AI/ML, 341 non-AI, and 221 not stated. Excluding 39 not-stated reports without an attributable AI component left a working pool of 182. Full-text term screening of the 182 reports yielded 14 hits (13/14 concordant; one classified as non-AI by arbitration) but no additional AI/ML devices. Rechecking 7 of the 54 AI/ML reports led to 6 exclusions and 1 retention (CSZ2300184), leaving 48 AI/ML devices.

Alt text: Two panels show the US FDA and China cohorts. Of 505 eligible US device families, 258 underwent dual coding and 200 entered the main analysis. The separate China analysis included 48 devices.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.docx