

Population representativeness cannot be assessed for most FDA-authorized AI medical devices targeting high-disparity conditions

Saptarshi Purkayastha

saptpurk@iu.edu

Indiana University Indianapolis

Aravind V. Kuruvikkattil

Indiana University Indianapolis

Chaitra Sree Sammeta

Indiana University Indianapolis

Aiswarya N. Perumbilly

Indiana University Indianapolis

Rashmita Kudamala

Indiana University Indianapolis

Judy W. Gichoya

Emory University

Article

Keywords:

Posted Date: October 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10808081/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Population representativeness cannot be assessed for most FDA-authorized AI medical devices targeting high-disparity conditions

Saptarshi Purkayastha^{**1}, Aravind V. Kuruvikkattil¹, Chaitra Sree Sammeta¹,
Aiswarya N. Perumbilly¹, Rashmita Kudamala¹, and Judy W. Gichoya²

¹Indiana University, Indianapolis, IN, USA

²Emory University, Atlanta, GA, USA

August 24, 2026

Abstract

Background Audits of FDA-authorized AI/ML medical devices ask whether submissions disclose demographic data. We ask three prior questions: for devices targeting conditions with documented demographic disparities, can training-data representativeness be assessed from the public record at all; does the underlying peer-reviewed evidence fill the gaps the summaries leave; and can equity constructs be coded reliably from these documents?

Methods We developed the 19-item Equity Readiness Assessment (ERA, development version) and a pre-registered Epidemiologic Crosswalk mapping each device’s target condition to its documented disparity profile, and applied them to all 1,428 devices with retrievable decision summaries through December 2025, using language-model extraction validated against dual independent, blinded human coding ($n = 153$). The nine items reproduced between human coders constitute the validated core, ERA-9, which carries all inferential claims.

Results Here we show that for 76.6% (95% confidence interval 73.8–79.3%) of the 899 devices targeting high-disparity conditions, representativeness cannot be assessed because race is unreported, a finding that rests on a reliably coded item and is near-invariant to how conditions are classified. Retrieving the pivotal publications rescues only 5 of 27 unassessable devices, locating the gap in the evidence base rather than the summary format. Graded equity judgements are not reproducible between trained coders (κ 0.01–0.12), whereas binary

^{**}Corresponding author: saptpurk@iu.edu

presence-of-reporting items are. No summary documents validation in a community or safety-net setting.

Conclusions Disclosure guidance has produced neither population-data alignment nor its assessability. We distil ERA-9 into a minimum equity-reporting standard for regulatory summaries and release the instrument, Crosswalk, and device-level dataset openly.

Plain language summary

Hundreds of medical devices that use artificial intelligence (AI) are approved in the United States each year, often for diseases that harm some groups of people more than others, such as stroke, heart disease, and several cancers. We asked a basic question: can anyone tell, from the public approval documents, whether these tools were tested on the people who carry most of the disease? For about three of every four AI devices aimed at such diseases, the answer is no, because the documents do not state the race or ethnicity of the people in the test data. The scientific papers behind the devices rarely fill this gap either. We also found that trained reviewers cannot reliably grade how well-matched a test population is when the documents give no numbers. We therefore propose a short list of facts that every approval summary should state, so that patients, clinicians, and hospitals can see who a device was tested on before it is used.

Introduction

Artificial-intelligence (AI) tools are now part of routine American health care: the U.S. Food and Drug Administration (FDA) has authorized 1,430 AI/ML-enabled medical devices through 30 December 2025, a more than thirty-fold increase over the past decade, and 1,428 have retrievable public decision summaries^{1;2}. Three-quarters read medical images in radiology; the rest span cardiology, neurology, ophthalmology, and other specialties. When tools of this reach carry systematic biases, individual errors become population-level inequities, a concern reflected in WHO ethics guidance³, in FDA’s health-equity priorities, and in a growing empirical literature on algorithmic bias in clinical AI^{4–6}.

A substantial body of work has audited what these documents disclose. Across independent audits using different instruments, race or ethnicity appears in roughly 4–24% of summaries depending on how generously a statement is counted, sex-specific results in under a third, subgroup performance in about one in eight, and socioeconomic status almost never^{7–12}. Adjacent audits have characterized what the devices do¹³, how completely benefit and risk are reported¹⁴, how often early recalls follow thin pre-market validation¹⁵, and whether clinical validation is reported at all¹⁶. The FDA’s own device centre has set out what AI/ML transparency should convey and to whom¹⁷, and Lui and Wang¹⁸ argue that these summaries can be technically dense while offering few entry points for fairness scrutiny. Supplementary Note 9 gives the audit-by-audit figures and reconciles them with ours.

These audits answer a transparency question (“did the device disclose its training data?”) but not an equity question (“do the disclosed data match the populations who need the device?”). A diabetic-retinopathy screening tool trained on a predominantly white academic cohort can score well on transparency while targeting a condition that burdens Black, Hispanic, and Native American populations two- to three-fold more heavily, and still be inequitable in deployment; Hassoon et al.¹⁹ articulate the underlying principle that representativeness against disease burden must run through every stage of AI development.

Existing standards bound the question from either side. The STANDING Together consensus recommendations, developed with more than 350 contributors across 58 countries, specify what a dataset custodian should disclose, but not a benchmark against which a disclosed composition should be judged, and they are not designed for retrospective third-party application to a regulatory document^{20;21}. The HEAL framework applies the same equity logic one stage downstream, asking whether measured *performance* is prioritized for the worst-off groups; it requires subgroup metrics that, as we show, almost no authorized device publishes²², and other frameworks operate at the organization or study level^{23–25}. Clinical-trial diversity work already

measures burden-adjusted representation with participation-to-prevalence ratios benchmarked against disease incidence rather than census shares^{26;27}. What is missing is the step in between: a benchmark for the development cohort itself, computable from the regulatory summary that clinicians, purchasers, and patients can actually see. That is what we build here.

The Equity Readiness Assessment (ERA) closes this gap. The ERA development version is a 19-item instrument across four domains: Demographic Reporting Depth, Population-Data Alignment, Equity-Relevant Design Features, and External Evidence Linkage. Its centrepiece maps each device’s target condition to a pre-registered Epidemiologic Crosswalk of disease-burden composition and disparity type, then asks whether the reported cohorts match the populations most affected. Because the equity meaning of under-representation differs by mechanism, the rubric branches by disparity type and scores detection-gap conditions, such as melanoma in dark skin^{28;29} or atrial fibrillation in Black patients, more strictly: there, an unrepresentative cohort perpetuates the very disparity the device should narrow. (We note a name-collision with an unrelated organizational-maturity “Equity Readiness Index”³⁰.)

This study is therefore built around three questions, of which the first is primary. First, for devices targeting a condition with a documented disparity, how often does the public record contain the demographic data needed to judge representativeness against burden at all? Assessability is logically prior to alignment: a cohort cannot be judged well- or ill-matched until its composition is stated. Second, is any gap an artefact of terse regulatory writing, or a property of the underlying evidence, answered by re-scoring a dual-source subset against the retrieved pivotal literature? Third, which equity constructs can be coded reliably from these documents in the first place, answered with a per-item reliability appraisal under dual blinded human coding; the nine items that two coders reproduced constitute the validated core of the instrument, which we designate ERA-9. Alongside these questions, we openly release the ERA instrument, the Epidemiologic Crosswalk, and the device-level dataset, and we test whether three FDA guidance milestones, the October 2021 Good Machine Learning Practice (GMLP) principles³¹, the December 2024 Predetermined Change Control Plan (PCCP) final guidance³², and the January 2025 lifecycle-management draft³³, are associated with detectable changes in reporting.

The answer to the third question is itself a finding. Several ERA items, including the graded alignment score, proved unreliable between trained coders. We report them descriptively, say so plainly, and anchor every inferential claim in this paper on the ERA-9 core.

Results

Device cohort and overall equity readiness

The cohort comprised 1,428 FDA-authorized AI/ML devices with retrievable decision summaries through 30 December 2025; two listed devices were excluded for FDA-side missing PDFs (Figure 1). Authorizations came through three routes: 510(k) clearance, the streamlined pathway based on equivalence to an earlier device (1,375 devices, 96.3%); De Novo, for novel low-to-moderate-risk devices (37, 2.6%); and Premarket Approval (PMA), the most stringent review (16, 1.1%). Radiology accounted for 1,094 of the 1,430 listed devices (76.5%; Supplementary Table S1). Document types differ markedly in length (510(k) median 10 pages; PMA median 33), so all scores are reported as the applicable-items-only ERA percentage and are stratified by pathway, never pooled.

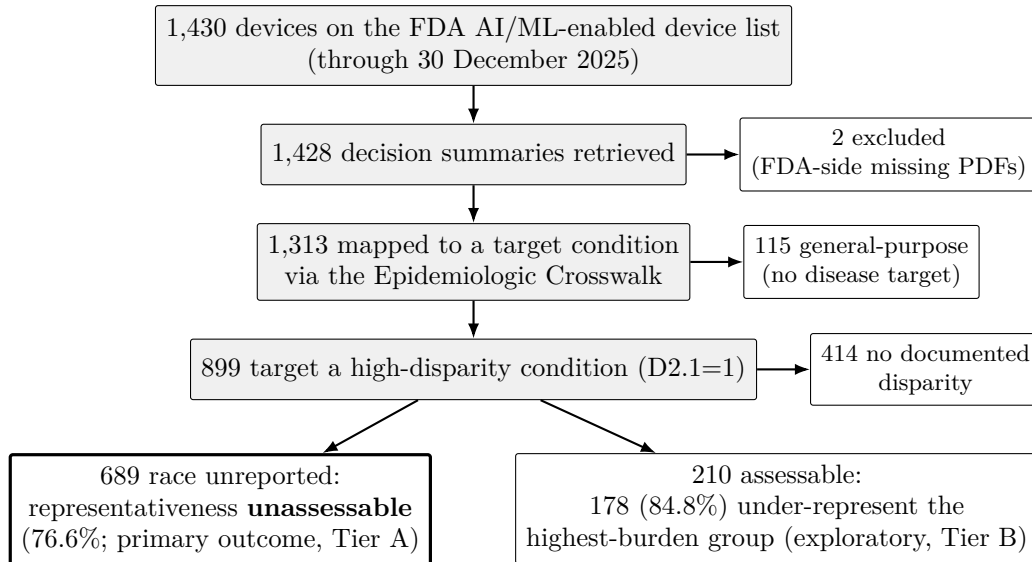


Figure 1: **Study flow.** Devices proceed from the FDA list through Crosswalk mapping to the primary analytic stratum. In parallel (not shown), a stratified $n = 153$ validation set was dual-coded by blinded human reviewers, and a dual-source arm re-scored the 68 devices with a retrievable pivotal publication (57 located).

Equity readiness was low and strongly pathway-dependent (Table 1): median applicable-items ERA was 12.5% for 510(k) devices (interquartile range [IQR] 4.2–29.2), 39.3% for De Novo (IQR 29.2–54.2), and 36.4% for PMA (IQR 34.3–45.5). The three-fold gap between 510(k) and the other pathways is the dominant structural signal and mirrors their documentation depth.

Which ERA items can bear inference

Because the composite ERA total was not reliably reproduced between independent coders (ICC(2,1) = 0.43; Supplementary Table S2), we do not treat it as an inferential quantity.

Instead, every coder-scored item was assigned, before any hypothesis was tested, to **Tier A** if the two blinded human reviewers reproduced it, meaning Cohen’s κ (a chance-corrected agreement statistic; 1 is perfect, 0 is chance) of at least 0.60 or its prevalence-adjusted equivalent for rare-positive items (Methods), and to **Tier B** otherwise.

Nine items met the Tier-A criterion: sex (D1.2), race/ethnicity (D1.3), geographic site (D1.5), socioeconomic status (D1.6), geographic diversity (D2.3), post-market demographic monitoring (D3.4), and all three external-evidence items (D4.1–D4.3). These nine constitute the validated core of the instrument, which we designate **ERA-9**. Seven items did not: age (D1.1), disease severity (D1.4), the graded representativeness score (D2.2), setting match (D2.4), and the three narrative Domain-3 items (D3.1–D3.3). The remaining three items are not tiered: D2.1 is Crosswalk-derived in the analysis rather than coder-assigned (Table 1, note), and D3.5 and D3.6 could not be assessed, because only one and no validation devices respectively were scored on them by both reviewers. The split is not arbitrary. Items asking whether a specific, named quantity appears in the document were reproduced; items requiring a coder to judge whether prose amounts to a comparison, a limitation, an exclusion, or a clinically appropriate setting were not.

All inferential claims in this paper, including the primary outcome, rest on ERA-9 items. Tier-B results are reported, because withholding them would misrepresent the instrument, but only as labelled descriptive statements.

Characteristic	510(k) (<i>n</i> = 1,375)	De Novo (<i>n</i> = 37)	PMA (<i>n</i> = 16)	Overall (<i>n</i> = 1,428)
Applicable-items ERA % (median, IQR) [§]	12.5 (4.2–29.2)	39.3 (29.2–54.2)	36.4 (34.3–45.5)	n/a
Race/ethnicity reported (D1.3 ≥ 1), % ^A	20.9	54.1	37.5	21.9
Targets high-disparity condition (D2.1=1), % ^C	62.0	83.8	100.0	63.0
Subgroup performance compared (D3.1), % ^B	15.1	30.2 (De Novo+PMA)		15.7

Table 1: **Cohort characteristics and equity readiness by regulatory pathway.** Applicable-items ERA percentage excludes conditional not-applicable items from both numerator and denominator; pooling across pathways is avoided because document-format heterogeneity would confound the comparison. ^ATier-A item (ERA-9 core; reproduced between blinded human coders; inference-grade). ^BTier-B item (not reproduced between coders; descriptive only). ^CCrosswalk-derived, and so outside the A/B scheme: the value in this row is computed deterministically by mapping the extracted target condition to the Epidemiologic Crosswalk, not assigned by a coder. The κ reported for D2.1 in Supplementary Table S2 belongs to a human quick-code that the analysis does not use. Its dependability is established empirically, by the near-invariance of the primary outcome to this classification (Results), and not by an inter-rater statistic. [§]The composite is a mixed-tier index with ICC(2,1) = 0.43 and is reported to describe the pathway gradient, not to support inference; the gradient itself is corroborated by the Tier-A race-reporting row.

Demographic reporting is sparse

Four Domain-1 axes met the Tier-A reliability criterion. Geographic site was reported by 43.6% of devices, sex by 39.7%, race or ethnicity by 21.9%, and socioeconomic status by only 0.2% (three devices) (Figure 2). Sex has a conditional not-applicable rule, so its rate is computed over the 1,321 sex-applicable devices; every other Domain-1 item applies to all 1,428. Performance actually stratified by an axis (score 2) was rarer still: 5.6% for race, 9.1% for sex. Race reporting was 2.6-fold higher in De Novo (54.1%) and 1.8-fold higher in PMA (37.5%) than in 510(k) (20.9%), again tracking document format rather than device design. Two further axes, age (56.2%) and disease severity (34.3%), are Tier-B and reported for completeness only.

Our race-reporting rate is higher than in some earlier audits because the ERA counts any race or ethnicity statement, including qualitative “diverse population” language; Supplementary Note 9 reconciles the rates across five independent audits, whose consistency underlines that it is the alignment question, addressed next, on which they are silent.

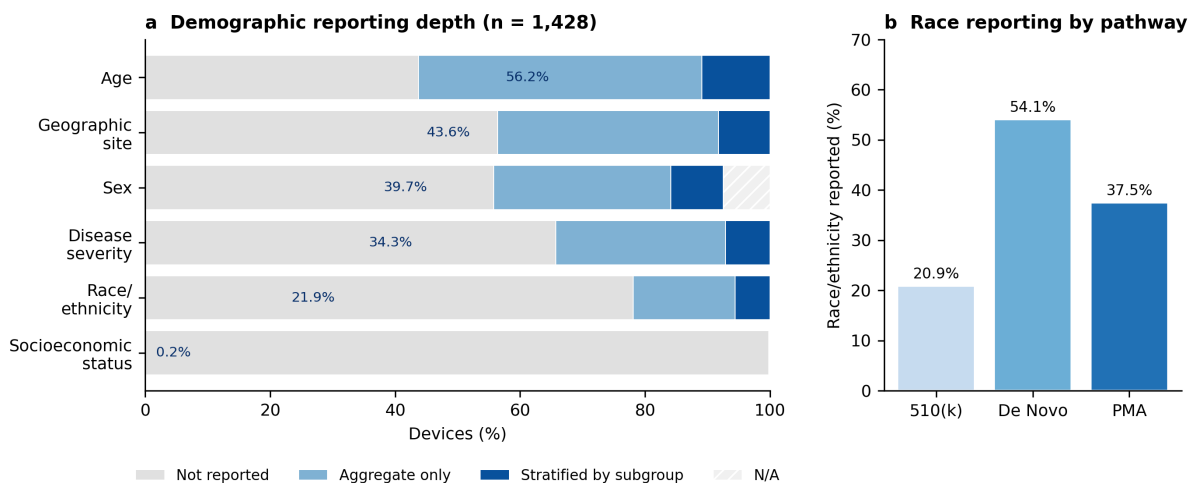


Figure 2: Demographic reporting depth, overall and by regulatory pathway. Socioeconomic status is essentially never reported (0.2%); performance stratified by any demographic axis (score 2) is rare across all items. In panel **a** the stacked segments are shares of all 1,428 devices, including the hatched not-applicable share that arises only for sex (single-sex-indicated devices); the percentage printed at each bar is the reported share (score 1 or 2) among *applicable* devices, which is the quantity given in the text.

Representativeness cannot be assessed for most high-disparity devices

Of the 1,313 condition-mapped devices (the remaining 115 were general-purpose tools with no disease target; Figure 1), 899 addressed a condition with a documented demographic disparity (D2.1=1). Among these 899 devices, 689 (76.6%; 95% Wilson confidence interval 73.8–79.3%) did not report race or ethnicity at all, so their training-data represen-

tativeness could not be evaluated against disease burden (Figure 3). This is the study's primary result and rests on an objective, reliably coded item (race reported: yes/no), not on any graded judgement.

These 899 devices cluster in the conditions that drive the most death and disability in the United States: intracranial haemorrhage and stroke (61 devices), coronary artery disease (38), lung cancer (36), atrial fibrillation (27), and breast, prostate, and colorectal cancer (Supplementary Table S8). Each carries substantially higher incidence or mortality in Black, Hispanic, or Native American populations (roughly 1.4- to 3-fold, depending on the condition), and together they affect tens of millions of Americans: the devices whose representativeness cannot be judged are aimed squarely at the nation's most inequitably distributed disease burden, the setting in which an unrepresentative training cohort matters most.

The primary result does not depend on how conditions were classified

Identifying the 899 devices requires classifying each target condition by disparity status, a step that was partly model-assisted (Methods; Supplementary Note 3) and therefore the most obvious place to look for fragility, so we bound its influence directly. If the classification were substantially wrong, the cannot-assess proportion inside the high-disparity stratum would diverge from the proportion outside it.

It does not. Race was unreported in 76.6% of the 899 high-disparity-condition devices, in 78.1% of the full cohort of 1,428, and, by subtraction, in roughly 80% of the 529 devices outside the stratum. Non-reporting is therefore close to uniform with respect to disparity status: even reclassifying every condition could move the headline proportion by no more than about four percentage points, and any plausible reclassification of a handful of conditions moves it far less: the stratum boundary is not doing the work.

The Crosswalk's contribution is therefore not to generate the proportion, which could be stated without it, but to establish that the devices concerned target the conditions where the omission carries the greatest consequence; the fuller version of this argument is in Supplementary Note 9.

For the 210 devices (23.4%) that did report race, we describe the representativeness scores produced by the branched Domain 2 rubric as an exploratory, Tier-B analysis. Severe under-representation (score 1) applied to 23.8% of these devices, moderate (score 2) to 61.0%, approximately proportional (score 3) to 12.9%, and over-represented or fully inclusive (score 4) to 2.4%; in total, 84.8% (95% CI 79.3–89.0%) scored ≤ 2 , indicating under-representation of the highest-burden group. Our pre-registered hypothesis that $\geq 60\%$ would score ≤ 1 was *not*

supported: under-representation clustered in the moderate band.

If we count the 689 unassessable devices together with the 50 that scored ≤ 1 , then 82.2% of all high-disparity-condition devices were either unevaluable or severely under-representative.

This gradient is Tier-B and we do not build on it. D2.2 is the least reliable item in the instrument (exact agreement 38%, $\kappa = 0.12$; Table 4). The 84.8% figure describes what the assessable devices look like and is stable under confidence stratification and alternative thresholds (Supplementary Table S4), but the conclusions of this paper follow from the Tier-A cannot-assess proportion and the dual-source test below, neither of which depends on D2.2.

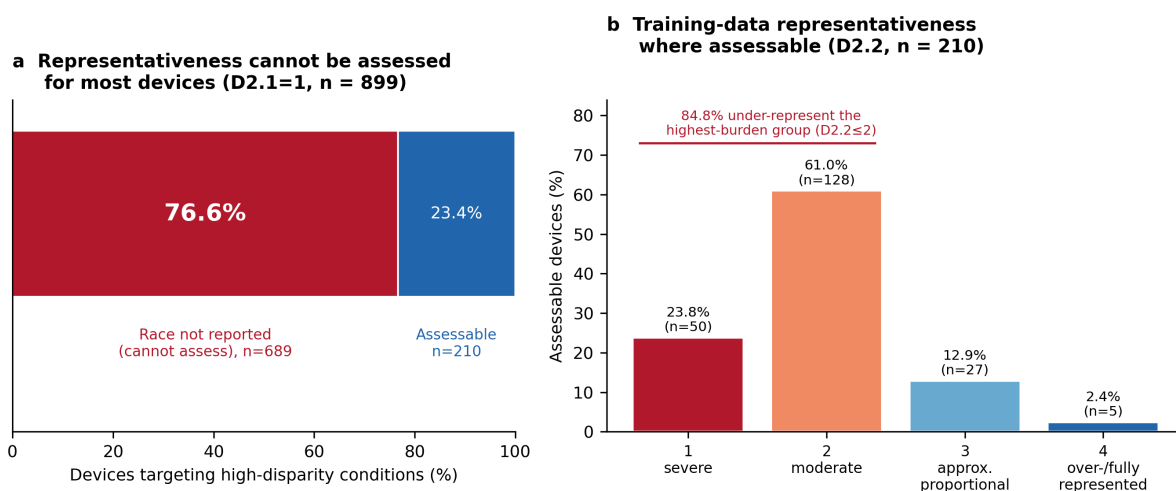


Figure 3: **The population-data alignment gap.** For most devices addressing conditions with documented disparities, the demographics required to judge representativeness are simply absent from the FDA decision summary; where present, reported cohorts under-represent the highest-burden group in the large majority of cases.

Equity-relevant design and external-evidence features are rare

Domain 3 and Domain 4 confirmed sparse lifecycle and evidence features (Table 2). Subgroup performance comparisons (D3.1) appeared in 15.7% of devices, population-limitation statements (D3.2) in 7.5%, and post-market subgroup monitoring (D3.4) in 0.8%. Among the 56 devices bearing a PCCP, demographic revalidation language (D3.5) appeared in 10 (17.9%). A referenced peer-reviewed publication (D4.1) was present in 5.2% and a summary-stated subgroup analysis (D4.3) in essentially none (0%). Every Domain 3–4 feature was several-fold more common in De Novo and PMA summaries than in 510(k) Summaries.

Validation rarely reaches the settings where gaps are widest

Two Domain-2 sub-items record *where* a device was validated. Geographic diversity (D2.3, Tier-A) asks whether validation spanned multiple, demographically varied regions; one device

Item	Overall %	510(k) %	De Novo+PMA %
D3.1 Subgroup performance comparison ^B	15.7	15.1	30.2
D3.2 Population-limitation statement ^B	7.5	6.1	43.4
D3.3 Exclusions specified ^B	40.1	38.3	86.8
D3.4 Post-market subgroup monitoring ^A	0.8	0.2	17.0
D3.5 PCCP demographic revalidation (of $n=56$ PCCP devices) [†]	17.9	18.9	0.0
D4.1 Peer-reviewed publication referenced ^A	5.2	4.5	22.6
D4.2 Summary-stated external validation ^A	0.4	0.2	3.8
D4.3 Summary-stated subgroup analysis ^A	0.0	0.0	0.0

Table 2: **Equity-relevant design and external-evidence features, overall and by pathway.** ^ATier-A (inference-grade); ^BTier-B (descriptive only). The three Tier-B items each require a coder to judge whether narrative prose constitutes a comparison, a limitation, or an exclusion, and independent reviewers did not converge on those judgements ($\kappa = 0.32, 0.01$ and 0.07 ; Supplementary Table S2); their rates are reported to describe the cohort, not to support a claim. [†]D3.5 is computed among the 56 devices carrying a Predetermined Change Control Plan (53 510(k), 3 De Novo, 0 PMA); 10 of 56 contain demographic revalidation language. Only one validation-set device had an applicable D3.5, so the item could not be tiered. D3.6 (adaptive subgroup drift monitoring) was positive in no device to which it applied, as anticipated for a still-emerging device category.

of 1,428 reached the top score (0.1%). Setting match (D2.4) asks whether validation occurred in a community, safety-net, or lower-resource setting appropriate to the condition’s care pathway; no device reached the top score, and 28.6% reached the partial score (Supplementary Table S7).

D2.4 is Tier-B, and its graded distribution should not be relied on ($\kappa = 0.02$). The ceiling result is a different and more robust claim, because it concerns a boundary rather than a gradient: neither human reviewer, nor the model, assigned the top anchor to any device, and no device in the cohort reached it. Across 1,428 authorizations, not one decision summary documents validation in a community, safety-net, or lower-resource setting appropriate to the target condition’s care pathway. AI tools are frequently promoted as a way to extend specialist-level capability to under-served settings, yet the authorized cohort was evaluated almost entirely in the well-resourced academic and referral environments that already have the most access.

Reporting rose modestly after the 2021 GMLP principles

We fitted a segmented (interrupted) time-series regression of applicable-items ERA percentage on authorization year, with pre-registered knots at the three guidance milestones and a specialty covariate. At GMLP, the model showed a level increase of +5.16 percentage points ($p = 0.001$) and a slope increase of +4.90 per year ($p < 0.001$) (Figure 4); the PCCP and lifecycle-draft knots showed no significant association ($p = 0.82, p = 0.35$), consistent with too few post-guidance authorizations, and ERA did not differ across specialty panels within 510(k) (Kruskal–Wallis $H = 4.51, p = 0.21$).

Three qualifications, detailed in Supplementary Note 9, apply: the model is fitted over ten

annual periods, so the later two knots are effectively unevaluable and the 2025 mean falls below the fitted trajectory; the outcome is the mixed-tier composite, so the trend tracks documentation depth broadly; and the August 2025 PCCP reissue³⁴ was not modelled as a separate breakpoint.

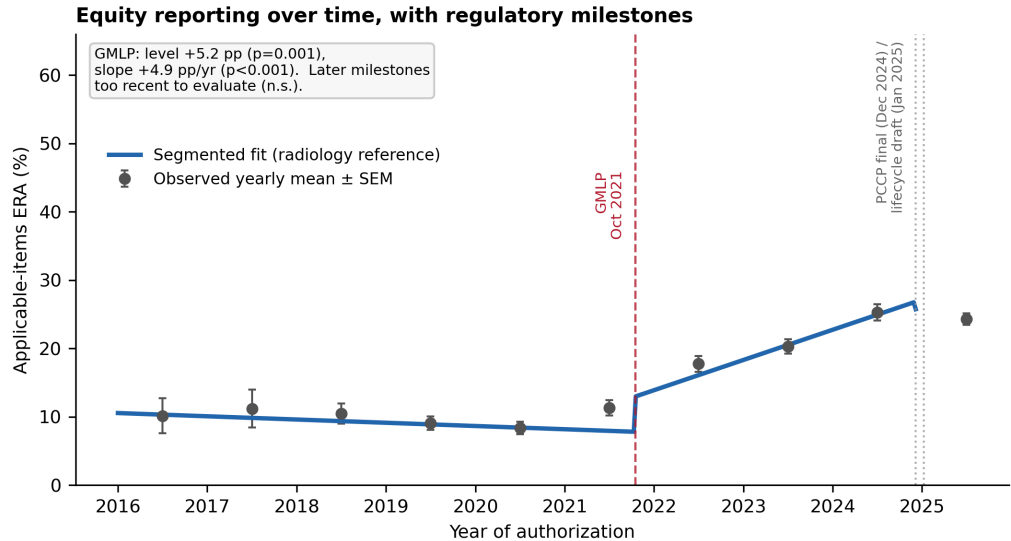


Figure 4: **Temporal trend with regulatory milestone breakpoints.** Equity reporting increased significantly after the 2021 GMLP principles; later guidance is too recent to evaluate.

Adding the primary literature does not close the gap

If the equity-data gap were merely an artefact of terse FDA writing conventions, it should shrink when the underlying peer-reviewed evidence is consulted. We therefore re-scored the 68 devices with a retrievable pivotal publication against the publication and the summary together (Methods). Publications were located for 57 of 68; adding them raised the applicable-items ERA percentage in 47 of 57 devices (82%), by a median of +8.7 percentage points, with the largest gains in 510(k) devices, whose summaries are thinnest.

The increase, however, was concentrated in items that do *not* bear on population-data alignment: the existence of a peer-reviewed publication itself (D4.1, upgraded in 34/57 devices), geographic sites (15/57), disease severity (11/57), sex (9/57), and publication-level subgroup or external-validation statements (10/57 and 9/57). Race reporting rose in only 5 of 57 devices, and socioeconomic status in only 1 (Supplementary Table S6). All 5 race additions occurred in devices that needed them.

Among the 57 located devices, 47 targeted a high-disparity condition (by numerical coincidence the same count as, though not the same set as, the 47 whose scores rose); 27 of these were unassessable under the decision summary, the publication rescued 5, and the remaining 22 stayed unassessable because their pivotal papers also omitted race. Rescue means only that

representativeness becomes assessable, not that it proves adequate: MelaFind is among the 5, and its publication reports a 97.9% White validation cohort for a detection-gap condition (Table 3).

The dual-source subset is better documented than the cohort as a whole, so this small test if anything overstates the literature’s contribution. Even so, the primary evidence closed the gap for fewer than one in five of the devices that needed it, indicating that the missing information is a property of the evidence base and not only of the FDA summary format.

Named devices show the gap is avoidable, not inevitable

The retrieved literature lets us contrast the cohort’s failures with its counter-examples at the level of individual, named devices (Table 3). At one extreme, the melanoma tools Nevisense and MelaFind report validation cohorts that are 97.5% and 97.9% White, for a cancer whose deadliest disparity is later-stage detection in patients with darker skin. At the other, the cervical-imaging device LUMA enrolled a cohort that was 31% African American and 21% Native American; the breast-mass device Imagio reported subgroup performance by race and age across 16 U.S. sites; and independent real-world validation of the Aidoc intracranial-haemorrhage tool covered 101,944 scans in a majority-Black cohort, with an intersectional subgroup analysis. Equitable evaluation is thus not precluded by device type, regulatory pathway, or imaging modality. It is simply uncommon.

Reliability, in full

Two reviewers independently coded a stratified $n = 153$ sub-sample (all 37 De Novo and 16 PMA summaries plus 100 stratified 510(k) devices), blinded to model output, with a third reviewer adjudicating disagreements to consensus. We report reliability for all 19 items (Supplementary Table S2), including those that performed poorly, because the distribution across items is itself a finding about what can be measured from these documents.

The instrument splits cleanly (Table 4). Nine items were reproduced between blinded coders and form Tier A, the ERA-9 core; the item that carries the primary outcome, whether race was reported, sits comfortably inside it (exact agreement 79.7%, $\kappa = 0.64$, model-versus-human κ 0.73–0.84). Seven tiered items were not reproduced, three of them at $\kappa < 0.10$, indistinguishable from chance agreement, and the graded representativeness score sits just above that floor at $\kappa = 0.12$. The composite ERA total has $\text{ICC}(2,1) = 0.43$, below any conventional threshold for a measurement instrument³⁵, which is why we tier at item level rather than reporting the composite as a score; full item-by-item statistics, including the prevalence-adjusted treatment

Device (indication)	Pathway	Validation cohort	Equity signal
<i>Under-representative exemplars</i>			
Nevisense (melanoma detection)	PMA	97.5% White; Fitzpatrick V-VI 1.5%	Detection-gap condition validated with near-absent dark-skin representation
MelaFind (melanoma detection)	PMA	97.9% White; 0.3% Black; no skin-tone reporting	Detection-gap condition, no dark-skin or skin-tone characterization
<i>Representative exemplars</i>			
LUMA (cervical neoplasia detection)	PMA	43% White, 31% African American, 21% Native American	Most demographically diverse cohort in the dual-source set
Imagio (breast-mass characterization)	PMA	16 U.S. sites; subgroup performance reported by race and age	Demographic subgroup analysis by race and age
Aidoc (intracranial-haemorrhage triage)	510(k)	101,944 scans; 56.5% Black (17-facility real-world validation)	Intersectional subgroup performance (independent external validation)

Table 3: **Worst- and best-case device exemplars from the dual-source set.** Cohort figures are drawn from the retrieved pivotal or independent-validation publications (Methods; Supplementary Information). Named counter-examples show that representative, subgroup-resolved validation is attainable across pathways and modalities.

of rare-positive items, are in Table 4 and Supplementary Table S2.

Seven tiered items were not reproduced. Three fell below $\kappa = 0.10$, which is indistinguishable from chance agreement: population-limitation statements ($\kappa = 0.01$), setting match ($\kappa = 0.02$), and specified exclusions ($\kappa = 0.07$); a fourth sub-0.10 statistic, the D2.1 human quick-code ($\kappa = 0.07$), belongs to a code the analysis does not use and is not tiered. The graded representativeness score sits just above that floor at $\kappa = 0.12$. The composite ERA total has ICC(2,1) = 0.43, which falls below any conventional threshold for a measurement instrument and is why we tier at item level rather than reporting the composite as a score.

We take the honest reading of this: the 19-item composite is not yet a reliable instrument, and we do not present it as one. What the study establishes is ERA-9 and a diagnosis of why the remaining items fail, namely that they ask a coder to interpret narrative prose instead of locating a stated quantity. That is tractable, and the deposited materials specify the rubric revisions it implies, but it is not solved here.

Model-versus-human agreement tracked human-versus-human agreement closely across items, indicating that the model reproduced the reviewers' behaviour, including where the reviewers disagreed with each other, rather than introducing an error mode of its own; the model's cohort-wide pathway medians also closely tracked the human-consensus medians on the validation set (Supplementary Table S8), supporting cohort-wide application of the extraction.

Tier	Item / statistic	Agreement	κ (R1 vs R2)	Model vs human κ
A	D1.3 Race/ethnicity reported (primary outcome)	79.7%	0.64	0.73–0.84
A	D3.4 Post-market subgroup monitoring	98.7%	0.87	0.81–0.94
A	D1.2 Sex	73.4%	0.73	0.76–0.81
A	D2.3 Geographic diversity	69.9%	0.71	0.58–0.65
A	D1.5 Geographic site	68.6%	0.65	0.68–0.80
A	D4.1 Publication referenced	88.2%	0.62	0.42–0.56
A	D1.6 Socioeconomic status [‡]	92.2%	0.32	0.23–0.66
A	D4.2 External validation stated [‡]	97.4%	0.32	0.49–0.66
A	D4.3 Subgroup analysis stated [¶]	100%	n.e.	n/a
B	D1.1 Age	49.7%	0.55	0.58–0.73
B	D1.4 Disease severity	22.2%	0.13	0.22–0.62
B	D3.1 Subgroup comparison	62.7%	0.32	0.28–0.51
B	D3.3 Exclusions specified	56.2%	0.07	–0.07–0.63
B	D2.4 Setting match	41.8%	0.02	–0.02–0.34
B	D3.2 Population limitations	25.7%	0.01	0.01–0.46
B	D2.2 Representativeness (graded)	38.0%	0.12	within ± 1 : 87%
—	ERA total (mixed tier), ICC(2,1)	n/a	0.43	n/a

Table 4: **Inter-rater and model-versus-human reliability by tier**, $n = 153$ **validation set**. Tier A items met the tier criterion (human–human $\kappa \geq 0.60$, or PABAK ≥ 0.75 where the kappa paradox applies), which was fixed before any hypothesis was tested, and constitute the ERA-9 core carrying all inferential claims; Tier B items did not and are reported descriptively. [‡]Kappa-paradox items: near-ceiling agreement with a rare positive class deflates κ , and PABAK (0.84 and 0.95 respectively) confirms reliability. [¶]D4.3: both reviewers scored every validation device negative, so agreement is perfect but Cohen’s κ is not estimable (n.e.) at zero prevalence; the item enters Tier A on PABAK = 1.00. D2.1 (human quick-code) and D3.5 are omitted here; D2.1 is computed from the Crosswalk rather than from coder judgement in the analysis, and D3.5 had one applicable validation device. The full 19-item table is Supplementary Table S2. The final row is ICC(2,1) for the composite, not Cohen’s κ ; it is reported as the diagnostic that motivated item-level tiering.

The instrument tracks independently verified equity evidence

We next asked whether ERA scores reflect genuine equity characteristics and not documentation length alone, comparing the decision-summary score against an independent ground truth taken from the retrieved primary literature: a known-groups design in which the label and the score come from separate documents. The 8 dual-source devices whose pivotal study independently reported race and conducted a demographic subgroup analysis scored markedly higher on the decision-summary ERA than the 38 whose publications reported neither (human-consensus median 50.9% versus 25.0%, one-sided Mann–Whitney $p = 0.003$; model-derived score, $p = 0.007$). Across all located devices, a count of equity evidence in the publication (range 0–4) correlated with the decision-summary ERA percentage (Spearman $\rho = 0.49$ human and 0.47 model, both $p < 0.001$), and a device that reported race in its FDA summary was more than five-fold likelier to report race in its independent publication (odds ratio 5.3, $p = 0.008$; Supplementary Table S9). Because the label and the score come from separate documents, this concordance indicates that ERA captures a real equity signal rather than submission verbosity.

Two caveats attach: the analysis was not pre-registered and supports construct validity rather than testing a hypothesis, and the score under test is the mixed-tier composite, although the item-level odds-ratio result, which uses only the Tier-A race item, does not inherit the composite’s reliability problem and points the same way.

Two named devices mark the boundary of what the score means: Nevisense scores above the median because its PMA summary reports demographics in full, even though its cohort is 97.5% White, while Aidoc scores low from its terse 510(k) summary despite an independent multi-site study documenting a majority-Black validation cohort. ERA measures equity readiness *as conveyed in the regulatory record*, not realized diversity: a device is flagged as unrepresentative only when it reports enough to be assessed, and a homogeneous cohort that reports nothing escapes scrutiny.

Discussion

Across a near-complete census of 1,428 FDA-authorized AI/ML medical devices, the answers to the study’s three questions point the same way. First, for three-quarters (76.6%) of devices targeting high-disparity conditions, the data needed to judge representativeness are absent: whether training data reflect the affected populations simply cannot be determined from the public record. Second, the gap is not an artefact of terse regulatory writing, because consulting the pivotal peer-reviewed studies rescued only 5 of 27 unassessable devices. Third, graded equity judgements cannot currently be coded reliably from these documents at all; only binary presence-of-reporting items, the ERA-9 core, were reproduced between trained coders. Where representativeness could be assessed, reported cohorts under-represented the highest-burden group in about 85% of cases, an exploratory Tier-B observation. Together these constitute a population-data alignment failure that disclosure requirements alone do not resolve.

Our pre-registered hypothesis, that a majority of high-disparity-condition devices would show *severe* under-representation, was not supported: where devices report enough to be scored, under-representation was more often moderate than severe. That does not weaken the equity concern; it relocates it upstream, to the near-absence of the demographic data required to compute any gradient at all.

The methodological finding deserves elaboration, because it constrains anyone who builds on this work. Graded alignment judgements made from regulatory prose are not currently reproducible between trained coders: our reviewers agreed on the graded representativeness score in 38% of cases and on whether a limitations section acknowledged a population gap in 26%. This is not a defect of our reviewers or rubric; it is what happens when a coder must

infer a quantitative comparison from documents that supply no quantities, and it predicts that any audit attempting graded equity scoring of these documents will meet the same ceiling. The remedy is source material that reports composition numerically, which our proposed reporting standard is designed to produce; until then, the defensible measurement is binary presence or absence, which is what the primary outcome uses.

The distinction between transparency and alignment matters because the two gaps have different remedies: a disclosure gap is closed by requiring a field, an alignment gap by requiring a benchmark. Prior audits measure what was disclosed^{7-9;13;14}; the ERA measures what was disclosed relative to what the epidemiology of the target condition requires, and neither STANDING Together nor HEAL is positioned to answer the question a procurement committee faces with a decision summary in hand: whether the cohort behind this authorization plausibly resembles the patients the device will see^{20;22}. Our contribution is to make that question computable from the public record, and to show where it currently cannot be answered.

The mismatch we document is not random: it tracks the structural determinants of clinical evidence generation, namely the geography of academic medical centers³⁶, the composition of research cohorts, and the economics of high-volume imaging, and predicate lineage plausibly propagates it, because a 510(k) device inherits its comparator from an earlier clearance whose evidence base often predates demographic-reporting expectations³⁷. Imaging models encode patient race even when humans cannot⁶, and such encodings surface as material under-diagnosis in under-served populations^{4;5}. The ERA quantifies, device by device, whether the upstream cohort gives a deployed model a reasonable chance of avoiding those failures, and the reliability distribution we report is a direct measurement of the interrogability that Lui and Wang¹⁸ predicted these documents would lack.

Regulatory guidance appears to have moved reporting but not alignment. The post-GMLP increase is encouraging, but GMLP and subsequent guidance encourage demographic *disclosure* rather than population-data *alignment*, and disclosure is precisely the layer at which the cohort still fails most often. One exception predates our breakpoints: FDA’s 2017 guidance on age-, race-, and ethnicity-specific device data already asks sponsors to enrol representative proportions consistent with the intended-use population, an alignment expectation³⁸. That guidance is non-binding and largely unmet, so the gap lies in enforcement and structure, not stated intent. Neither the PCCP final guidance^{32;34} nor the lifecycle draft³³ imposes a representativeness criterion, and model cards³⁹ and the ONC HTI-1 transparency rule⁴⁰ likewise govern disclosure. FDA’s own account of AI/ML transparency is framed around communicating information to the right audiences¹⁷, which is necessary but, on this evidence, insufficient: information that

was never generated cannot be communicated.

One development since our cutoff bears on these recommendations. In January 2026 FDA revised its Clinical Decision Support Software guidance, widening enforcement discretion under the statutory non-device exclusion⁴¹. Where it applies, it shrinks the public record this study audits; but image- and signal-analysis software remains regulated, so the devices whose representativeness we find unassessable stay inside the premarket pathway. If anything, the boundary strengthens the case for structured fields: equity-relevant information about non-device tools will now often exist nowhere public, and the tools leaving premarket review are the ones purchasers are least able to evaluate independently.

Methodologically, we add an explicit reliability floor to an increasingly accepted practice. LLM-assisted evidence extraction is used where properly validated^{42–44}; we did not develop a new extraction method, but pair temperature-0 extraction with blinded dual human validation, three-rater agreement metrics, and inference anchored on the items that humans and the model code concordantly. All prompts, schema, code, and item-level data are released⁴⁵ to enable replication on EU MDR, UK MHRA, Health Canada, and PMDA cohorts.

This work has limitations, and the first is the most important. **The ERA is not yet a reliable 19-item instrument.** Seven of the sixteen tiered items were not reproduced between blinded coders, three of them at $\kappa < 0.10$, and the composite has $\text{ICC}(2,1) = 0.43$, below any accepted threshold for a measurement instrument³⁵. We responded by restricting inference to ERA-9, and a reader should take from this that ERA-9, not a thirty-point score, is what has been validated; adopters should take the core and treat the full 19-item set as a development version and a research agenda. The deposited materials record which rubric anchors failed and how, so that the next version can build on the diagnosis instead of repeating it.

Second, the graded representativeness score (D2.2), the item that most directly expresses the study’s central construct, is also its least reliable ($\kappa = 0.12$). We treat it descriptively, but the demotion has a cost: we can show that alignment is unevaluable far more convincingly than we can show how badly the evaluable devices align. The trial-diversity literature measures this construct with continuous ratios such as the participation-to-prevalence ratio^{26;27}; those require cohort composition as percentages, which FDA summaries rarely report, so an ordinal rubric was the only option at cohort scale. Where percentages are recoverable, a ratio-based measure would be preferable.

Third, coding-team capacity was finite: the Crosswalk classification of novel conditions and the graded D2.2 scoring were performed by the language model rather than by dual human review, both documented as deviations⁴⁵. The model reproduced the 13 human anchor classifi-

423 cations 13/13 in isolation and 11/13 interleaved, a thin base we do not present as sufficient on its
 424 own; the substantive defence is the primary outcome’s near-invariance to the classification (Re-
 425 sults) and the deposited amendment log, which exposes every call to inspection. An expanded
 426 human re-classification of a random sample of the 746 model-classified conditions is the first
 427 thing we would add with further capacity. More generally, the study observes a strict claims
 428 hierarchy with respect to model involvement. Inferential claims rest only on ERA-9 items, for
 429 which model-versus-human agreement matched or exceeded human-versus-human agreement.
 430 Quantities that depend on model coding without complete human review, namely the Crosswalk
 431 expansion and the graded D2.2 scores, support only the descriptive and exploratory statements
 432 labelled as such, and the primary outcome is protected twice over, by its Tier-A status and by
 433 its near-invariance to the Crosswalk classification. The pre-registered escalation rule that would
 434 have expanded human review for low-agreement items was not executed for the Tier-B items
 435 (Methods); demoting those items from inference is the substitute remedy, not an equivalent
 436 one, and completing that review is the extension we would prioritize.

437 Fourth, the ERA measures equity readiness *as conveyed in public regulatory documents*,
 438 which is what regulators, purchasers, and clinicians actually see, not confidential submission
 439 contents or non-public manufacturer practice. Fifth, the Crosswalk is U.S.-centric (GBD, CDC,
 440 and ACS estimates) and the cohort is U.S.-regulatory, so both generalize to other jurisdic-
 441 tions only with adaptation. Sixth, this is a reporting audit, not a performance audit: as the
 442 Nevisense–Aidoc contrast shows, the record and the reality can diverge, a high score signals
 443 assessable, well-characterized reporting rather than a demonstrably representative cohort, and
 444 even fairness-tuned models can lose subgroup performance under real-world distribution shift⁴⁶.
 445 Seventh, the FDA list grows continuously; our snapshot closes at 30 December 2025 and should
 446 be read as a complete census of a defined window, with the open dataset designed to make
 447 re-audit at a later cutoff straightforward.

448 The implications differ by audience. *For regulators*, the alignment gap is addressable through
 449 a minimum equity-reporting standard for AI/ML device submissions. Reporting guidelines
 450 already govern AI clinical studies^{47–49}, and STANDING Together sets the consensus expectation
 451 for documenting datasets²⁰, but none is tailored to the short regulatory summary that is often
 452 the only document a purchaser or clinician will read. We therefore distil ERA-9 into eight
 453 structured fields that the FDA could require in decision summaries (Table 5). Each field restates
 454 a STANDING Together recommendation or an expectation already present in FDA’s own 2017
 455 demographic guidance³⁸, where it is at present advisory; our contribution is the mapping,
 456 the mandate, and the auditability, and the fields are, by construction, the ones our reliability

analysis shows can be coded consistently once reported as values rather than prose. The GMLP experience shows explicit expectations measurably move reporting. A statutory vehicle exists but is stalled: the 2022 Food and Drug Omnibus Reform Act requires Diversity Action Plans for certain device studies, yet the implementing draft guidance⁵⁰, removed from the FDA website in January 2025 and restored by court order the following month, had not been finalized as of mid-2026, so attaching structured fields to the AI lifecycle draft³³ is the more robust near-term route.

For health systems and clinicians, the practical message is that the decision summary usually cannot tell them whom a device was tested on. ERA-style assessment belongs in procurement alongside performance and cost review, and the released device-level dataset makes that immediately possible. *For manufacturers*, the named counter-examples show that representative, subgroup-resolved validation is attainable across pathways and modalities; the eight fields specify exactly what a submission would need to state, and in a market where equity documentation is rare, demonstrating it is a differentiator rather than a burden. *For patients and the public*, the finding is simply stated: for most AI devices aimed at the diseases that burden minority communities hardest, no one outside the company can currently verify who the device was tested on. *For researchers*, the open dataset provides a baseline for longitudinal surveillance and cross-jurisdiction replication, and the reliability diagnosis specifies which equity constructs the next instrument can and cannot expect to code from prose.

Methods

Study design and data source

This cross-sectional audit covers all AI/ML-enabled medical devices on the FDA’s public AI/ML-Enabled Device List¹, cutoff 30 December 2025, list snapshot downloaded 30 April 2026 (1,430 listed; 1,428 with retrievable decision-summary PDFs from `accessdata.fda.gov`; two logged as FDA-side missing). The unit of analysis is the device authorization: 1,375 510(k) Summaries, 37 De Novo Decision Summaries, and 16 PMA Summaries of Safety and Effectiveness. During quality control, 26 De Novo cases were initially retrieved as the short Classification Order letter. We corrected these to the full Decision Summary and re-extracted them. Both human reviewers had coded all De Novo cases from the Decision Summaries throughout, so human reliability is unaffected (deviations log,⁴⁵). Seven image-only PDFs (1995–2003) were processed with `ocrmypdf` before extraction.

Field	What a submission would report	ERA item(s)
Target condition and disparity axis	Condition and the demographic group(s) it disproportionately burdens, with source	D2.1
Cohort race/ethnicity	Race/ethnicity distribution of training and validation cohorts	D1.3
Cohort age, sex, and skin tone	Age and sex distribution; Fitzpatrick skin type where clinically relevant	D1.1, D1.2
Socioeconomic and access context	Payer/insurance mix or area-level socioeconomic status of the cohort	D1.6
Sites and care setting	Number and type of sites (academic, community, safety-net) and geography	D1.5, D2.3, D2.4
Population-data alignment	Explicit comparison of cohort composition to the condition's disease-burden composition	D2.2
Subgroup performance	Performance disaggregated by race, sex, and age (and skin tone), with confidence intervals	D3.1, D4.3
Post-market subgroup monitoring	Plan to monitor subgroup performance after authorization	D3.4

Table 5: **A proposed minimum equity-reporting standard for AI/ML device decision summaries.** Eight structured fields distilled from the ERA, positioned within the CONSORT-AI/DECIDE-AI/TRIPOD+AI reporting-guideline family and as a regulatory-summary projection of the STANDING Together consensus recommendations²⁰, but tailored to the short document that procurement committees and clinicians actually read. Each maps to an ERA item so that compliance is directly auditable. Every field requires a reported value in place of a narrative statement, which is the property our reliability analysis identifies as the precondition for consistent third-party coding.

The ERA instrument

The ERA (version 2.7, locked for deposit) comprises 19 items across four domains. Domain 1 (Demographic Reporting Depth) has six items: age, sex, race/ethnicity, severity, geographic site, and socioeconomic status. Each is scored 0 (absent), 1 (aggregate), or 2 (subgroup-stratified), and sex can be marked not-applicable for single-sex-indicated devices. The domain maximum is 12. Domain 2 (Population-Data Alignment) has four scored items plus a classifier. These are D2.1 condition burden asymmetry (binary), a disparity-type classifier (incidence, mortality, detection-gap, access, reference-standard, or mixed), D2.2 representativeness (0–4, with the rubric branching by disparity type and detection-gap scored strictly), and geographic and setting diversity (each 0/1/2). The domain maximum is 9. Domain 3 (Equity-Relevant Design) has six binary items: subgroup comparison, limitations, exclusions, post-market monitoring, PCCP demographic revalidation, and adaptive drift monitoring. The domain maximum is 6. Domain 4 (External Evidence) has three binary items, which cover referenced publications and summary-stated external validation and subgroup analysis. The domain maximum is 3. The reported score is the applicable-items-only percentage. It is the points obtained divided by the maximum possible over applicable items, with not-applicable items excluded from both the numerator and the denominator. The nine items whose human–human reliability met the

Tier-A criterion (Statistical analysis) constitute the validated core, ERA-9; the full 19-item set is the development version, and the two are distinguished throughout. Full scoring rules are deposited⁴⁵.

Epidemiologic Crosswalk

Each target condition was mapped to its documented demographic-burden profile and disparity type. We used CDC WONDER⁵¹, GBD 2021⁵², ACS Cancer Statistics⁵³, and condition-specific literature, and we recorded the highest-burden group(s), the disparity type, and the reference. Thirteen anchor conditions were classified by human reviewers before scoring began and retain their human classification throughout; the remaining 746 conditions encountered during extraction were classified by the model under the deposited expansion procedure (Supplementary Note 3). Domain 2 items were not extracted directly. Instead, the model extracted the target condition and demographic composition, and D2.1, its type, and D2.2 were computed post hoc by mapping to the Crosswalk and applying the branched rubric. D2.2 is not applicable when race is unreported (D1.3=0) or the condition has no documented disparity (D2.1=0).

Crosswalk classification was validated by re-classifying the 13 human anchors under production conditions, with 11 of 13 reproduced. We state plainly that 13 anchors is a thin validation base for a step that defines the primary stratum, and we have therefore bounded that step's influence on the primary outcome empirically, and do not rest on the anchor concordance alone (Results). Two properties of the classification limit how far an error could propagate. It operates on conditions, not devices, so it is invariant to anything in an individual submission; and it is a binary presence-of-disparity call for the primary outcome, with the finer disparity-type distinction entering only the Tier-B D2.2 rubric. The deposited amendment log records every added condition with its evidence and adjudication, so any individual classification can be checked and contested by a reader⁴⁵.

LLM-assisted extraction and classification

In accordance with journal policy on the use of large language models, we document all model-assisted steps here. Extraction used Claude Sonnet 4.6 (Anthropic) via the Anthropic API⁵⁴. Each decision summary was processed as a PDF, with deposited system and user prompts and a structured output schema⁴⁵. Extraction ran at temperature 0, a logged deviation from the pre-registered configuration (Supplementary Note 2); temperature 0 constrains but does not eliminate run-to-run variation in a served model, and we claim no bitwise determinism.

Human coding-team capacity was exhausted after the validation sub-sample. Two classifica-

tion steps that had been planned for dual human review were therefore performed by the model instead, and each was validated against the human-coded gold standard. The first was classification of novel Crosswalk conditions by disparity type. The second was graded D2.2 scoring of the 210 assessable devices, validated against the human D2.2 consensus (87% within ± 1). Both steps feed Tier-B quantities only, and neither enters the primary outcome except through the definition of the high-disparity stratum, whose influence we bound directly in Results. A general-purpose pipeline for extracting structured fields from FDA decision summaries with a large language model has been described and benchmarked independently⁴⁴; our extraction was developed separately, and what we add is the equity instrument, not the extraction method. The model did not author any text of this manuscript. Every departure from the pre-registered protocol, model-assisted or otherwise, is enumerated with its rationale in Supplementary Note 7 and in the deposited deviations register⁴⁵.

Human validation and consensus

A stratified sub-sample of $n = 153$ ($\approx 11\%$ of the cohort) was drawn with a fixed random seed⁴⁵. It was a census of all 37 De Novo and 16 PMA summaries, plus 100 510(k) Summaries allocated across year-band $\times$ specialty strata. Two reviewers (A.N.P., C.S.S.) independently coded all 153 devices, blinded to model output. A third reviewer (S.P.) adjudicated all disagreements and produced a consensus item matrix, which served as the human gold standard. Primary coding used the FDA decision summary only, to preserve cross-pathway comparability. We then carried out a parallel dual-source scoring for the 68 devices with a retrievable second source.

Redesign of the second-source arm (deviation). The pre-registered protocol specified public labeling and Instructions for Use as the second source for a triangulation sub-sample of $n = 50$. Before the arm was coded, we substituted the primary literature and expanded the frame to the 68 devices with a retrievable second source, because concordance within FDA paperwork could not answer whether the gap exists in the underlying science. The cost is that the pre-specified labeling-concordance estimate is not reported; the full rationale is in Supplementary Note 8 and the deposited register⁴⁵.

For each device in this arm, we added the retrieved primary study to the source set and upgraded a consensus code wherever the publication reported content that the summary had omitted. The upgrade followed a deterministic mapping from the study’s extracted demographic, multi-site, and subgroup-analysis fields to the ERA anchors. The mapping is conservative and defaults to the lower anchor when a subgroup axis or site count is not explicitly resolvable, so the reported Arm-1-to-Arm-2 delta is a lower bound. Condition burden asymmetry (D2.1)

in both arms is the Epidemiologic Crosswalk value, pre-coded by condition as the instrument specifies, which is the same source used for the cohort analysis. Domain-4 items under this arm are the pre-registered “verified” constructs, coded from the publication in hand. The decision-summary-only versus dual-source delta (Supplementary Information) quantifies whether the equity-data gap reflects FDA writing conventions or the primary evidence itself.

Statistical analysis

Reliability tiering. Before hypothesis testing, each coder-scored item (16 of the 19: D2.1 is Crosswalk-derived in the analysis, and D3.5–D3.6 lacked applicable validation devices) was assigned to Tier A or Tier B on a single criterion: Tier A required human–human Cohen’s $\kappa \geq 0.60$, or, where exact agreement reached 90% with a rare positive class so that the kappa paradox applies, $\text{PABAK} \geq 0.75$ ⁵⁵. Inference is restricted to Tier A. This tiering was introduced after the pre-registered analysis plan was written, in response to the observed reliability distribution, and is logged as a deviation⁴⁵; it is a restriction on the claims we make, not a selection among results, and reliability for all items is reported regardless of tier. The pre-registered per-item escalation rule (expand human review where model-versus-human $\kappa < 0.70$) was evaluated on the axis in which it was specified. It was not triggered for the primary-outcome item (model–human κ 0.73–0.84). It would have been triggered for several Tier-B items, and coding capacity to execute the 40% expansion was not available; tiering those items out of inference is the substitute remedy, and we record it as such rather than presenting the escalation rule as satisfied.

Change of primary outcome (deviation). The pre-registered primary outcome was the D2.2 distribution among devices with D2.1=1; that analysis is reported in full, and its hypothesis was not supported. Because D2.2 proved to be the least reliable item in the instrument ($\kappa = 0.12$), we judged that it should not carry the study’s primary claim and designated as primary a quantity pre-specified within the same analysis and stratum: the proportion of D2.1=1 devices for which D2.2 is inapplicable because race is unreported. The change was made after reliability was known and before the dual-source and validity analyses were run; the full rationale is in Supplementary Note 8 and the deposited register⁴⁵.

The primary outcome as reported was therefore the proportion of high-disparity-condition devices (D2.1=1) for which representativeness could not be assessed because race was unreported. We report it with a 95% Wilson confidence interval. As a robustness check on the stratum definition, specified when the outcome was redesignated and before the comparison was run, we compare this proportion with race non-reporting across the full cohort. It was not part of the pre-registered plan. Secondary analyses covered several measures. These were

the D2.2 distribution among assessable devices (with Wilson intervals) and the pre-registered hypothesis test ($\geq 60\%$ scoring ≤ 1); item-level frequencies; the applicable-items ERA percentage by pathway (median and IQR); a Kruskal–Wallis comparison across specialty panels within pathway; and a segmented (interrupted) time-series regression of ERA percentage on authorization year. The regression used knots at October 2021 (GMLP), December 2024 (PCCP final), and January 2025 (lifecycle-management draft), together with a specialty covariate. Reliability was reported per item in several ways: percent agreement, Cohen’s κ ⁵⁶ (quadratic-weighted for ordinal items), prevalence-and-bias-adjusted κ ⁵⁵ for skewed items, and ICC(2,1) for the ERA total³⁵. Model-versus-human agreement was assessed in a three-rater design. A pre-registered adaptive-escalation rule governed any item with $\kappa < 0.70$. As a post-hoc criterion-validity check, we tested whether decision-summary ERA scores discriminated devices that the independent primary literature showed to be equity-attentive (known-groups design), using the Mann–Whitney U test for the dichotomous contrast, Spearman correlation for the continuous equity-evidence count, and Fisher’s exact test for item-level concordance. This analysis was not pre-registered and is reported as supporting construct validity. Analyses used Python (pandas, scipy, statsmodels). The study follows PRISMA-ScR⁵⁷, GATHER⁵⁸, and STROBE⁵⁹.

Deviations from the pre-registered protocol. Nine departures occurred. Each is enumerated with its rationale in Supplementary Note 7 (register) and Supplementary Note 8 (full prose for the most consequential three: the change of primary outcome, the redesign of the second-source arm, and the substitution of model coding for planned dual human review of the Crosswalk expansion and D2.2, with the escalation rule not executed for Tier-B items), and logged in the deposited register⁴⁵.

Ethics

The study involved no human participants and analyzed only publicly available regulatory documents. An Indiana University Institutional Review Board not-human-subjects determination is on file.

Data availability

The complete ERA dataset (item-level scores for all 1,428 devices, metadata, consensus codes, and Crosswalk classifications) is deposited with the OSF project and will receive a versioned DOI at acceptance. The pre-registration (instrument v2.7, coding workbook, Crosswalk, prompts, schema, sampling design and seed, adaptive-escalation rule, deviations log, and analysis plan) is at [10.17605/OSF.IO/QSGMV](https://doi.org/10.17605/OSF.IO/QSGMV) (osf.io/b34zf).

Code availability

All extraction, classification, scoring, and analysis code is released under the MIT license at the OSF project⁴⁵; the instrument, workbook, and Crosswalk are released under CC-BY 4.0.

Acknowledgements

The authors acknowledge the Open Science Framework (OSF) for providing the platform used to preregister the study and archive the study materials. We also acknowledge the U.S. Food and Drug Administration for providing public access to the regulatory documents analyzed in this study.

Author contributions

S.P. conceived the study, developed the instrument and pipeline, and drafted the manuscript. A.V.K. contributed to instrument refinement and Crosswalk construction. C.S.S. and A.N.P. independently coded the $n = 153$ validation set; S.P. adjudicated. R.K. contributed to Crosswalk construction and figures. J.W.G. provided AI-fairness and regulatory expertise, contributed to instrument design, and supervised. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Funding

The authors declare no specific funding for this work.

References

- [1] U.S. Food and Drug Administration. Artificial intelligence-enabled medical devices. U.S. Food and Drug Administration. Formerly titled “Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices”, 2025. URL <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices>. Accessed 24 August 2026.

- [2] Stan Benjamens, Pranavsingh Dhunoo, and Bertalan Meskó. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digital Medicine*, 3:118, 2020. doi: 10.1038/s41746-020-00324-0.
- [3] World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance, 2021. URL <https://www.who.int/publications/i/item/9789240029200>.
- [4] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019. doi: 10.1126/science.aax2342.
- [5] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021. doi: 10.1038/s41591-021-01595-0.
- [6] Judy W. Gichoya, Imon Banerjee, Ananth R. Bhimireddy, John L. Burns, Leo A. Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P. Lungren, Lyle J. Palmer, Brandon J. Price, Saptarshi Purkayastha, Ayis T. Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, and Haoran Zhang. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022. doi: 10.1016/S2589-7500(22)00063-2.
- [7] Vijaytha Muralidharan, Boluwatife Adeleye Adewale, Caroline J. Huang, Mfon Thelma Nta, et al. A scoping review of reporting gaps in FDA-approved AI medical devices. *npj Digital Medicine*, 7:273, 2024. doi: 10.1038/s41746-024-01270-x.
- [8] Viraj Mehta et al. Evaluating transparency in AI/ML model characteristics for FDA-reviewed medical devices. *npj Digital Medicine*, 8:673, 2025. doi: 10.1038/s41746-025-02052-9.
- [9] Daniel Windecker, Giovanni Baj, Isaac Shiri, Pooya Mohammadi Kazaj, Johannes Kaesmacher, Christoph Gräni, and George C. M. Siontis. Generalizability of FDA-approved AI-enabled medical devices for clinical use. *JAMA Network Open*, 8(4):e258052, 2025. doi: 10.1001/jamanetworkopen.2025.8052.

- [10] Ram Sivakumar, Brian Lue, and Shinjini Kundu. FDA approval of artificial intelligence and machine learning devices in radiology: a systematic review. *JAMA Network Open*, 8(11):e2542338, 2025. doi: 10.1001/jamanetworkopen.2025.42338.
- [11] Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho, and James Zou. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nature Medicine*, 27(4):582–584, 2021. doi: 10.1038/s41591-021-01312-x.
- [12] Bassel Almarie, Luis Fernando Gonzalez-Gonzalez, Lucas Antônio dos Santos Barbosa, Amelie Lutz, Ulrich Grosse, and Felipe Fregni. Machine learning-enabled medical devices authorized by the US Food and Drug Administration in 2024: regulatory characteristics, predicate lineage, and transparency reporting. *Biomedicines*, 13(12):3005, 2025. doi: 10.3390/biomedicines13123005.
- [13] Rohan Singh, Monika Bapna, Abdul Rahman Diab, Emily S. Ruiz, and William Lotter. How AI is used in FDA-authorized medical devices: a taxonomy across 1,016 authorizations. *npj Digital Medicine*, 8:388, 2025. doi: 10.1038/s41746-025-01800-1.
- [14] John C. Lin, Bhav Jain, Jay M. Iyer, Ishan Rola, Anusha R. Srinivasan, Chaerim Kang, et al. Benefit-risk reporting for FDA-cleared artificial intelligence-enabled medical devices. *JAMA Health Forum*, 6(9):e253351, 2025. doi: 10.1001/jamahealthforum.2025.3351.
- [15] Branden Lee, Patrick Kramer, Sara Sandri, Ritika Chanda, Crystal Favorito, Olivia Nasef, et al. Early recalls and clinical validation gaps in artificial intelligence-enabled medical devices. *JAMA Health Forum*, 6(8):e253172, 2025. doi: 10.1001/jamahealthforum.2025.3172.
- [16] Sammy Chouffani El Fassi, Adonis Abdullah, Ying Fang, Sarabesh Natarajan, Awab Bin Masroor, Naya Kayali, Simran Prakash, and Gail E. Henderson. Not all AI health tools with regulatory authorization are clinically validated. *Nature Medicine*, 30(10):2718–2720, 2024. doi: 10.1038/s41591-024-03203-3.
- [17] Aubrey A. Shick, Christina M. Webber, Nooshin Kiarashi, Jessica P. Weinberg, Aneesh Deoras, Nicholas Petrick, Anindita Saha, and Matthew C. Diamond. Transparency of artificial intelligence/machine learning-enabled medical devices. *npj Digital Medicine*, 7:21, 2024. doi: 10.1038/s41746-023-00992-8.
- [18] Chun-Wang Lui and Pusheng Wang. The fairness paradox in public regulatory summaries of FDA-authorized artificial intelligence-enabled medical devices. *Frontiers in Medicine*, 13:1894905, 2026. doi: 10.3389/fmed.2026.1894905.

- [19] Ahmed Hassoon, Christine Lin, Hyun Yi Woo, Ruxandra Irimia, Jill A. Marsteller, Anthony Li, Antonio Banderia, Hubert Leo, Xiaoyi Peng, David Rastall, and Mark Dredze. Guiding artificial intelligence in public health and medicine with epidemiology: a lifecycle framework for mitigating AI misalignment. *Annals of Epidemiology*, 112:119–126, 2025. doi: 10.1016/j.annepidem.2025.11.004.
- [20] Joseph E. Alderman, Joanne Palmer, Elinor Laws, Melissa D. McCradden, Johan Ordish, Marzyeh Ghassemi, Stephen R. Pfohl, Negar Rostamzadeh, Heather Cole-Lewis, Ben Glocker, Melanie Calvert, Tom J. Pollard, Lauren Oakden-Rayner, Judy W. Gichoya, Alastair K. Denniston, Xiaoxuan Liu, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. *The Lancet Digital Health*, 7(1):e64–e88, 2025. doi: 10.1016/S2589-7500(24)00224-3.
- [21] Anmol Arora, Joseph E. Alderman, Joanne Palmer, Shaswath Ganapathi, Elinor Laws, Melissa D. McCradden, Lauren Oakden-Rayner, Stephen R. Pfohl, Marzyeh Ghassemi, Francis McKay, Darren Treanor, Negar Rostamzadeh, Bilal Mateen, Jacqui Gath, Adewale O. Adebajo, Stephanie Kuku, Rubeta Matin, Katherine Heller, Elizabeth Sapey, Neil J. Sebire, Heather Cole-Lewis, Melanie Calvert, Alastair Denniston, and Xiaoxuan Liu. The value of standards for health datasets in artificial intelligence-based applications. *Nature Medicine*, 29(11):2929–2938, 2023. doi: 10.1038/s41591-023-02608-w.
- [22] Mike Schaeckermann, Terry Spitz, Malcolm Pyles, Heather Cole-Lewis, Ellery Wulczyn, Stephen R. Pfohl, Donald Martin, Jr., Ronnachai Jaroensri, Geoff Keeling, Yuan Liu, Jenna Lester, Yun Liu, Ivor Horn, Po-Hsuan Cameron Chen, et al. Health equity assessment of machine learning performance (HEAL): a framework and dermatology AI model case study. *eClinicalMedicine*, 70:102479, 2024. doi: 10.1016/j.eclinm.2024.102479.
- [23] Jee Young Kim et al. Development and preliminary testing of health equity across the AI lifecycle (HEAAL): a framework for healthcare delivery organizations to mitigate the risk of AI solutions worsening health inequities. *PLOS Digital Health*, 3(5):e0000390, 2024. doi: 10.1371/journal.pdig.0000390.
- [24] Jacob G. Ellen, João Matos, Martin Viola, Jack Gallifant, Justin Quion, Leo Anthony Celi, et al. Participant flow diagrams for health equity in AI. *Journal of Biomedical Informatics*, 152:104631, 2024. doi: 10.1016/j.jbi.2024.104631.
- [25] Mingxuan Liu, Yilin Ning, Salinelat Teixayavong, Xiaoxuan Liu, Mayli Mertens, Marcus

Eng Hock Ong, Nan Liu, et al. A scoping review and evidence-gap analysis of clinical AI fairness. *npj Digital Medicine*, 8:360, 2025. doi: 10.1038/s41746-025-01667-2.

[26] Tanvee Varma, Michelle Mello, Joseph S. Ross, Cary P. Gross, and Jennifer E. Miller. Metrics, baseline scores, and a tool to improve sponsor performance on clinical trial diversity: retrospective cross sectional study. *BMJ Medicine*, 2(1):e000395, 2023. doi: 10.1136/bmjmed-2022-000395.

[27] Megan Y. Gimmen, Kekoa Taparra, Curtiland Deville, and Scarlett Lin Gomez. Promoting equitable patient representation in therapeutic clinical trials by accounting for population disease burden. *Scientific Reports*, 15(1):10591, 2025. doi: 10.1038/s41598-025-91174-x.

[28] Adewole S. Adamson and Avery Smith. Machine learning and health care disparities in dermatology. *JAMA Dermatology*, 154(11):1247–1248, 2018. doi: 10.1001/jamadermatol.2018.2348.

[29] Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A. C. Allerup, Utako Okata-Karigane, James Zou, and Albert S. Chiou. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(32):eabq6147, 2022. doi: 10.1126/sciadv.abq6147.

[30] Laura Wagon, Marcia Nißen, and Tobias Kowatsch. Equity readiness index for digital health organizations. medRxiv, 2025.

[31] U.S. Food and Drug Administration, Health Canada, and U.K. Medicines and Healthcare products Regulatory Agency. Good machine learning practice for medical device development: Guiding principles, October 2021. URL <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles>.

[32] U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Guidance for industry and FDA staff. Final guidance issued 4 December 2024. Docket FDA-2022-D-2628. Superseded by the reissued version of 18 August 2025, December 2024. URL <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intel>

- [33] U.S. Food and Drug Administration. Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations: Draft guidance. Draft guidance issued 7 January 2025. Docket FDA-2024-D-4488, January 2025. URL <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence-enabled-device-software-functions>
- [34] U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Guidance for industry and FDA staff (reissued). Reissued final guidance, 18 August 2025. Docket FDA-2022-D-2628, August 2025. URL <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence-enabled-device-software-functions>
- [35] Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012.
- [36] Amit Kaushal, Russ Altman, and Curt Langlotz. Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA*, 324(12):1212–1213, 2020. doi: 10.1001/jama.2020.12067.
- [37] Urs J. Muehlematter, Christian Bluethgen, and Kerstin N. Vokinger. FDA-cleared artificial intelligence and machine learning-based medical devices and their 510(k) predicate networks. *The Lancet Digital Health*, 5(9):e618–e626, 2023. doi: 10.1016/S2589-7500(23)00126-7.
- [38] U.S. Food and Drug Administration. Evaluation and reporting of age-, race-, and ethnicity-specific data in medical device clinical studies: Guidance for industry and food and drug administration staff. Final guidance, issued 12 September 2017. Docket FDA-2016-D-0734, 2017. URL <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/evaluation-and-reporting-age-race-and-ethnicity-specific-data-medical-device-clinical-studies>
- [39] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* ’19)*, pages 220–229, 2019. doi: 10.1145/3287560.3287596.

- [40] U.S. Office of the National Coordinator for Health Information Technology. Health data, technology, and interoperability: Certification program updates, algorithm transparency, and information sharing (HTI-1); final rule. 89 FR 1192, 9 January 2024, FR Doc. 2023-28857, 45 CFR Parts 170 and 171, 2024. URL <https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-tra>
- [41] U.S. Food and Drug Administration. Clinical decision support software: Guidance for industry and food and drug administration staff. Revised final guidance issued 6 January 2026 and re-issued 29 January 2026, superseding the final guidance of 28 September 2022. Docket FDA-2017-D-6569, January 2026. URL <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software>.
- [42] Qusai Khraisha, Sophie Put, Johanna Kappenberg, Azza Warraitch, and Kristin Hadfield. Can large language models replace humans in systematic reviews? evaluating GPT-4’s efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*, 15(4):616–626, 2024. doi: 10.1002/jrsm.1715.
- [43] Gerald Gartlehner, Leila Kahwati, Rainer Hilscher, Ian Thomas, Shannon Kugley, Karen Crotty, Meera Viswanathan, Barbara Nussbaumer-Streit, Graham Booth, Nathaniel Erskine, Amanda Konet, and Robert Chew. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. *Research Synthesis Methods*, 15(4): 576–589, 2024. doi: 10.1002/jrsm.1710.
- [44] Hanyang Li, Xiao He, Adarsh Subbaswamy, Patrick Vossler, Alexej Gossmann, Karandeep Singh, and Jean Feng. Scaling medical device regulatory science using large language models. *npj Digital Medicine*, 9(1):221, 2026. doi: 10.1038/s41746-026-02353-7.
- [45] Saptarshi Purkayastha, Aravind V. Kuruvikkattil, Chaitra Sree Sammeta, Aiswarya N. Perumbilly, Rashmita Kudamala, and Judy W. Gichoya. Beyond transparency: equity readiness of FDA-authorized AI/ML-enabled medical devices. OSF Pre-Registration, 2026. URL <https://osf.io/qsgmv/>.
- [46] Yuzhe Yang, Haoran Zhang, Judy W. Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, 30(10): 2838–2848, 2024. doi: 10.1038/s41591-024-03113-4.

- [47] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, Alastair K. Denniston, and the SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, 26(9):1364–1374, 2020. doi: 10.1038/s41591-020-1034-x.
- [48] Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A. Clifton, Gary S. Collins, Spiros Denaxas, Alastair K. Denniston, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28(5):924–933, 2022. doi: 10.1038/s41591-022-01772-9.
- [49] Gary S. Collins, Karel G. M. Moons, Paula Dhiman, Richard D. Riley, Andrew L. Beam, Ben Van Calster, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.
- [50] U.S. Food and Drug Administration. Diversity action plans to improve enrollment of participants from underrepresented populations in clinical studies: Draft guidance for industry, June 2024. URL <https://www.federalregister.gov/documents/2024/06/28/2024-14284/diversity-action-plans-to-improve-enrollment-of-participants-from-underrepresented-populations>
- [51] Centers for Disease Control and Prevention. CDC WONDER: Online databases, 2025. URL <https://wonder.cdc.gov/>. Accessed 24 August 2026.
- [52] GBD 2021 Diseases and Injuries Collaborators. Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet*, 403(10440):2133–2161, 2024. doi: 10.1016/S0140-6736(24)00757-8.
- [53] American Cancer Society. Cancer facts & figures 2024, 2024. URL <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures.html>.
- [54] Anthropic. System card: Claude Sonnet 4.6. Anthropic model system card, published 17 February 2026, February 2026. URL <https://www.anthropic.com/claude-sonnet-4-6-system-card>.
- [55] Ted Byrt, Janet Bishop, and John B. Carlin. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423–429, 1993. doi: 10.1016/0895-4356(93)90018-V.

- 876 [56] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for cate-
877 gorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.
- 878 [57] Andrea C. Tricco, Erin Lillie, Wasifa Zarín, Kelly K. O’Brien, Heather Colquhoun, Danielle
879 Levac, David Moher, Micah D. J. Peters, Tanya Horsley, Laura Weeks, Susanne Hempel,
880 et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation.
881 *Annals of Internal Medicine*, 169(7):467–473, 2018. doi: 10.7326/M18-0850.
- 882 [58] Gretchen A. Stevens, Leontine Alkema, Robert E. Black, J. Ties Boerma, Gary S. Collins,
883 Majid Ezzati, et al. Guidelines for accurate and transparent health estimates reporting: the
884 GATHER statement. *The Lancet*, 388(10062):e19–e23, 2016. doi: 10.1016/S0140-6736(16)
885 30388-9.
- 886 [59] Erik von Elm, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche,
887 Jan P. Vandenbroucke, and for the STROBE Initiative. The strengthening the reporting
888 of observational studies in epidemiology (STROBE) statement: guidelines for reporting
889 observational studies. *Annals of Internal Medicine*, 147(8):573–577, 2007. doi: 10.7326/
890 0003-4819-147-8-200710160-00010.

Supplementary Files

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

- [Supplementary.pdf](#)