

Supplementary Information

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

Contents

1	Supplementary Note 1: Cohort assembly and document correction	2
2	Supplementary Note 2: LLM-assisted extraction	2
3	Supplementary Note 3: LLM-assisted Epidemiologic Crosswalk classification (protocol deviation)	3
4	Supplementary Note 4: LLM-assisted D2.2 representativeness scoring (protocol deviation)	4
5	Supplementary Note 5: Consensus construction and data-quality control	5
6	Supplementary Note 6: Dual-source triangulation: demographic reporting in the primary literature	5
7	Supplementary Note 7: Complete register of deviations from pre-registration	6
8	Supplementary Note 8: Full prose rationales for the principal protocol deviations	8
9	Supplementary Note 9: Prior transparency audits, reconciliation of rates, and additional analytic qualifications	8
	Supplementary Tables (S1–S10)	9
	Supplementary references	17

List of Tables

S1	Cohort by FDA specialty panel	10
S2	Per-item inter-rater and model–human reliability by tier	11
S3	Epidemiologic Crosswalk disparity-type distribution	12
S4	D2.2 confidence-stratified sensitivity	12
S5	Demographic reporting: FDA decision summaries versus the primary literature	12
S6	Item-level upgrades from dual-sourcing (Arm 1 to Arm 2)	13

S7	Cross-validation of cohort-wide extraction against the human-adjudicated subset	13
S8	Frequently targeted high-disparity conditions, with U.S. disease burden	14
S9	Criterion and convergent validity of the ERA against the primary literature . .	15
S10	Complete register of deviations from the pre-registration	16

1 Supplementary Note 1: Cohort assembly and document correction

The audited cohort was the FDA public AI/ML-Enabled Device List [1], cutoff 30 December 2025: 1,430 listed submissions, of which 1,428 had retrievable decision-summary PDFs from `accessdata.fda.gov`; two (one 510(k), one PMA supplement) were logged as FDA-side missing (`download_failures.csv`). Document types were 1,375 510(k) Summaries, 37 De Novo Decision Summaries, and 16 PMA Summaries of Safety and Effectiveness (SSEDs). During quality control, 26 of the 37 De Novo cases had initially been retrieved as the short FDA Classification Order letter instead of the detailed Decision Summary specified by the protocol. We re-downloaded these as Decision Summaries and re-extracted them under the identical pipeline. Because both human reviewers coded all De Novo cases from the Decision Summaries throughout, human reliability is unaffected. Seven image-only PDFs (1995–2003) were processed with `ocrmypdf` (Tesseract backend) before extraction, and OCR fidelity was spot-checked. The specialty-panel distribution is given in Table S1.

2 Supplementary Note 2: LLM-assisted extraction

Extraction used Claude Sonnet 4.6 (Anthropic) via the Anthropic API [2]. Each decision summary was supplied as a PDF attachment with a fixed study-level system prompt and a per-device user prompt (both deposited, 3). The output schema constrained each ERA item to an integer-or-null score plus an `applicable` flag. Responses were parsed and then validated against the schema post hoc. The full corpus was extracted a single time. The Domain 2 alignment items (D2.1, D2.1_type, D2.2) were not extracted directly. Instead, the model extracted the target condition and the reported demographic composition, and these items were computed post hoc (Supplementary Notes 3–4).

Sampling parameters (protocol deviation). The deposited instrument (v2.7, Part D) specified that temperature would not be set, with reproducibility resting on a fixed model version and a fixed prompt. Production extraction was run at temperature 0. The two descriptions are not interchangeable and the discrepancy is logged in the deposited deviations register [3]. We note the distinction because reproducibility claims should be checkable against the deposit, and because temperature 0 is often over-read: it constrains the sampling distribution but does not guarantee identical output across runs of a served model. Anthropic’s own documentation is explicit on this point: model weights are fixed for a given model ID, but the surrounding serving infrastructure, including the request router, safety classifiers and sampling logic, can change

over time [4]. The same documentation confirms that `claude-sonnet-4-6` is a pinned snapshot from the 4.6 generation onward and not an evergreen alias, so the model version underlying our extraction is fixed and identifiable. Where we observed run-to-run divergence (Supplementary Note 3), we attribute it to context composition and residual serving non-determinism, not to sampling temperature, and we make no bitwise-determinism claim anywhere in this work. Reproducibility here rests on the deposited prompts, schema, model version, code, and raw outputs, which together permit an independent re-run and a comparison of results, not on the assumption that a re-run returns identical tokens.

3 Supplementary Note 3: LLM-assisted Epidemiologic Crosswalk classification (protocol deviation)

The pre-registered plan called for two human reviewers to classify each distinct target condition in the Epidemiologic Crosswalk by disparity type. After the reviewers completed the $n = 153$ validation set, coding-team capacity was exhausted; the remaining 746 novel conditions were therefore classified by the language model (Sonnet 4.6, temperature 0, forced structured output), retaining the 13 human-seeded “known” conditions as fixed anchors. This deviation and its validation are recorded in the deposited deviations log [3].

Validation. We re-classified the 13 human-seeded conditions with the model. It reproduced the human disparity type in 13/13 when the seeds were classified in isolation, and in 11/13 (85%) when they were interleaved with novel conditions. The two divergences were borderline sub-type calls: incidental pulmonary embolism was assigned *none* on one run, and melanoma was assigned *mixed* rather than *detection-gap*. The relevant variable is the interleaving, not the sampling temperature: the divergences appeared only in the condition where surrounding novel conditions changed the classification context, and both are sub-type boundary calls of exactly the kind that context would be expected to shift. Because “known” conditions retain their human classification in the analysis, these divergences do not propagate.

Adequacy of this validation. Thirteen anchors is a thin base for validating a step that defines the primary analytic stratum, and we do not present it as sufficient on its own. Three further considerations bound the risk. First, the classification operates on conditions rather than devices, so it cannot be influenced by anything in an individual submission and cannot introduce device-level differential error. Second, the primary outcome depends only on the binary presence-of-disparity call, not on the finer disparity-type distinction, and the sub-type distinction is where both observed divergences occurred; disparity type enters only the Tier-B D2.2 rubric. Third, and most directly, the primary outcome is empirically near-invariant to this classification: race was unreported in 76.6% of the 899 high-disparity devices, in 78.1% of all 1,428, and in roughly 80% of the 529 devices outside the stratum, so even wholesale reclassification could move the headline by no more than about four percentage points, and plausible reclassification of individual conditions far less (main text). A human re-classification of a random sample of the 746 model-classified conditions would nonetheless materially strengthen the Crosswalk as a reusable artefact, and is recorded in the deposit as the highest-priority exten-

sion. We note that melanoma is the anchor condition for the detection-gap branch of the D2.2 rubric and for the skin-cancer exemplars in the main text. Its classification in the analysis is the human one, so the exemplar analysis does not depend on the model’s call. We would still not run disparity-type classification unsupervised. Device-weighted, 899 of 1,313 condition-mapped devices (68.5%) mapped to a condition with a documented disparity (D2.1=1). The disparity-type distribution is in Table S3. All high-device-count *none* classifications were general-purpose imaging or measurement tools with no specific disease target, consistent with human expectation.

4 Supplementary Note 4: LLM-assisted D2.2 representativeness scoring (protocol deviation)

The graded Domain 2 representativeness item (D2.2, 0–4) requires comparing a device’s reported training-data demographic composition against the disease-burden composition of the target condition. Two constraints shaped how we scored it. The extraction stored race/ethnicity as an ordinal reporting-depth score and not as percentages, and human coding capacity was exhausted. D2.2 for the 210 assessable devices (D2.1=1 and race reported) was therefore scored by the language model. The model applied the Instrument Part E rubric to each device’s extracted demographic evidence, branched by the crosswalk disparity type. D2.2 is not applicable, and was not scored, when race was unreported (D1.3=0) or the condition had no documented disparity (D2.1=0).

Validation. Against the human-adjudicated D2.2 consensus on the overlapping validation devices ($n = 23$), the model matched exactly in 52% of cases and to within ± 1 point in 87%. This is comparable to the human–human agreement on this item, which was the least reliable in the instrument (38% exact, $\kappa = 0.12$; Table S2). The model attached a confidence level to each score (35 high, 73 medium, and 102 low), and low confidence corresponded to qualitative rather than quantitative demographic evidence. The proportion scoring $D2.2 \leq 2$ was 84.8% overall, 72.2% among high and medium-confidence scores, and 91.4% among high-confidence scores (Table S4). The pattern is not monotonic, which complicates a simple reading. Low confidence marks devices whose demographic evidence was qualitative, and a qualitative statement such as “diverse population” cannot establish proportionality against a burden profile, so the rubric will tend to place such devices at or below a score of 2. Part of the overall 84.8% therefore reflects the scoring rule as much as measured cohort composition. The conclusion holds within the high-confidence stratum, where quantitative composition was available, and we regard that stratum as the more defensible estimate. Because of this item’s intrinsic unreliability, the primary analysis is anchored on the objective “cannot assess” proportion (main text), and D2.2 is reported descriptively.

5 Supplementary Note 5: Consensus construction and data-quality control

The human gold standard is a consensus matrix over all 153 devices $\times$ 19 items. Each cell took one of three values. Where the reviewers matched, it took their agreed code. Where they disagreed, it took the adjudicated code (808 cells). Where the instrument’s deterministic not-applicable rule applied, it took not-applicable (for example, D2.2 is N/A when D2.1=0 or D1.3=0; 127 cells). Two data-quality issues were corrected before analysis and are logged [3]. The first was 40 out-of-range entries by one reviewer on binary items: 39 values of “2” on D3.1/D3.2 were mapped to 1, and one “o”-for-“0” typo on D1.3 was fixed. The second was 43 cells (across 32 devices) where the two reviewers disagreed on *applicability*, with one coding not-applicable and the other coding a value. The disagreement worklist had excluded these cells from adjudication, so we resolved them to the single available code and flagged them. The ERA-total intraclass correlation was computed by summing the literal item scores, not by reading the spreadsheet’s cached formula cells.

6 Supplementary Note 6: Dual-source triangulation: demographic reporting in the primary literature

For a dual-source subset of 68 devices, we retrieved the primary clinical evidence and coded what that literature itself reports. These were devices flagged for triangulation or with an identifiable pivotal publication, and the evidence retrieved was peer-reviewed pivotal studies or FDA-cited publications. The aim was to test whether the demographic gap is an artifact of FDA-summary writing conventions or is present in the underlying evidence. Primary publications were located for 57 of 68 devices. Even in this peer-reviewed literature, demographic reporting was thin (Table S5). Race or ethnicity was reported in only 16/57 (28%) and socioeconomic status in 2/57 (4%), with a demographic subgroup performance analysis in 11/57 (19%). Where race was reported for skin or breast imaging devices, cohorts were frequently 95–98% of a single group.

Paired within-device delta (Arm 1 vs Arm 2). To quantify the effect of dual-sourcing on the ERA score itself, we re-scored each of the 57 located devices with the primary study added to the source set (Arm 2). We upgraded a consensus code wherever the publication reported content that the decision summary had omitted. The mapping from the study’s extracted fields to the ERA anchors is deterministic and conservative. It defaults to the lower anchor when a subgroup axis or site count is not explicitly resolvable, so the reported delta is a lower bound. Several items were carried from Arm 1 unchanged: the Domain-2 alignment metadata (D2.1), the graded representativeness value (D2.2), the Domain-3 device and design items, and setting match (D2.4, which requires clinical-pathway judgement). Adding the literature raised the applicable-items ERA percentage in 47 of 57 devices (82%), from a mean of 33.6% to 41.7% (a mean paired change of +8.0 and a median of +8.7; the change is computed on unrounded per-device values, so it differs from the difference of the rounded endpoint means). The largest

gains were in 510(k) devices, whose summaries are thinnest (mean +10.9, versus De Novo +7.8 and PMA +4.8). The increase was concentrated in items orthogonal to population-data alignment (Table S6). These were the existence of a peer-reviewed publication (D4.1), geographic sites (D1.5), disease severity (D1.4), sex (D1.2), and publication-level subgroup and external-validation statements (D4.2/D4.3). Race reporting rose in only 5 of 57 devices, and socioeconomic status in only 1. Among the 57 located devices, 47 targeted a high-disparity condition (D2.1=1; by numerical coincidence the same count as, though not the same set as, the 47 devices whose scores rose), and 27 of these were unassessable for representativeness under the decision summary because race was unreported. Adding the publication rescued 5 of those 27 (K190072, K230223, K250096, P090012, P200003), and the other 22 remained unassessable because their pivotal papers also omitted race. D2.1 in this arm is the Epidemiologic Crosswalk value, pre-coded by condition as the instrument specifies, which is the same source used for the cohort analysis in the main text. The equity-evidence gap is therefore largely, though not wholly, a property of the primary evidence and not only of the FDA paperwork. Dual-sourcing raises age, sex, severity, geography, and external-evidence codes, but it barely moves race or socioeconomic status, and it leaves more than four fifths of the high-disparity devices that need a representativeness assessment still without one.

7 Supplementary Note 7: Complete register of deviations from pre-registration

The instrument, analysis plan, sampling design and Crosswalk were deposited before coding began [3]. Nine departures from that protocol occurred. We enumerate all of them here, including those that are routine quality-control corrections, so that the deposit and the manuscript can be read against each other without reconciliation. Each is also recorded, with dates and supporting files, in the deposited register (`deviations_log.csv`). Table S10 classifies each by whether it could plausibly affect the primary outcome.

Three deviations warrant comment beyond the table.

D6, change of primary outcome. This is the most consequential. The protocol named the D2.2 distribution among D2.1=1 devices as primary. D2.2 proved to be the least reliable item in the instrument, and we judged that a study should not rest its principal claim on an item two trained coders reproduced 38% of the time. We promoted a quantity from within the same pre-registered analysis, on the same device stratum: the proportion of that stratum for which D2.2 is inapplicable because race is unreported. This was already specified in the protocol as D2.2’s N/A condition, so no new construct was introduced; what changed was which quantity carries the headline. The original specification is reported unchanged in the main text, including the pre-registered hypothesis ($\geq 60\%$ scoring ≤ 1), which was not supported. The decision was taken after reliability was computed and before the dual-source and known-groups analyses were run.

D7, adaptive escalation not executed. The protocol committed to expanding human review to 40% of the cohort (≈ 571 devices) for any item failing the κ threshold. Evaluated

on the model-versus-human axis in which it was specified, the rule was not triggered for the primary-outcome item (κ 0.73–0.84). It would have been triggered for several Tier-B items, and the coding capacity to execute a 571-device expansion was not available. We did not execute it, and we do not claim to have satisfied it. The substituted remedy is reliability tiering (D8): rather than expanding review to rescue unreliable items, we removed them from inference. This is a weaker remedy than the one promised, and it is the clearest respect in which this study falls short of its protocol.

D8, reliability tiering. Introduced after the reliability distribution was observed. Tiering is a restriction on the claims made, not a selection among results: reliability for all 19 items is reported regardless of tier (Table S2), the tier criterion is a single pre-stated threshold applied uniformly, and no item was moved between tiers after the analyses were run. Readers can therefore reconstruct any analysis at any tier boundary they prefer.

8 Supplementary Note 8: Full prose rationales for the principal protocol deviations

This note preserves, at full length, the rationales for the two deviations that most affect interpretation; the complete nine-entry register is Supplementary Note 7 and Table S10. The sampling-parameters deviation is documented in Supplementary Note 2, and the substitution of model coding for planned dual human review is documented in Supplementary Notes 3–4.

Change of primary outcome. The pre-registered plan named the Domain 2 score and the D2.2 distribution among devices with D2.1=1 as the primary outcome. We report that analysis in full in the main text, including the pre-registered hypothesis test, which was not supported. However, D2.2 proved to be the least reliable item in the instrument ($\kappa = 0.12$), and we judged that an unreliable item should not carry a study’s primary claim. We therefore designated as primary a different quantity drawn from the same pre-registered analysis and the same device stratum: the proportion of D2.1=1 devices for which D2.2 is inapplicable because race is unreported. This quantity was pre-specified in the instrument as the N/A rule governing D2.2 and was always a component of the Domain 2 analysis; what changed is its promotion from a denominator condition to the headline. The change was made after reliability was known and before the dual-source and validity analyses were run, and it is recorded in the deposited deviations register [3]. Readers who prefer the original specification will find it reported unchanged in the main text; its hypothesis was not supported, and the study’s conclusions do not rest on it.

Redesign of the second-source arm. The pre-registered protocol specified a triangulation sub-sample of $n = 50$ (16 PMA, 17 De Novo, 17 510(k)) in which the second source was public labeling and Instructions for Use. We changed both the sample and the comparator document class. Labeling and IFU are drawn from the same regulatory apparatus as the decision summary, so concordance between them would have tested how far equity content migrates within FDA paperwork. The sharper question is whether the gap exists in the underlying science or only in the paperwork, and answering it requires the peer-reviewed evidence, not the labeling. We therefore substituted the primary literature as the second source and expanded the frame to the 68 devices for which a second source was retrievable, comprising devices flagged for triangulation plus those with an identifiable pivotal publication. The change was made before the arm was coded and is recorded in the deposited register [3]. The cost is that the per-pathway labeling-concordance estimate specified in the protocol is not reported; the benefit is a direct test of whether the equity-data gap is an artefact of regulatory writing conventions, which the labeling comparison could not have delivered.

9 Supplementary Note 9: Prior transparency audits, reconciliation of rates, and additional analytic qualifications

Audit-by-audit figures. Muralidharan et al. [5] reviewed 692 decision summaries and found race or ethnicity reported in only 3.6% and socioeconomic status in 0.9%. Mehta et al. [6]

developed a 17-item transparency score (mean 3.3) across more than 1,000 devices, with any dataset demographics reported in 23.7%. Windecker et al. [7] found sex-specific results in 28.7% and age subgroups in 23.2% of the 505 devices (of 903) that reported a clinical performance study, and Sivakumar et al. [8] showed that fewer than 30% of 723 radiology devices had clinical testing. Foundational work by Wu et al. [9] reported that only 13% of devices considered subgroup performance, and Almarie et al. [10] replicated the disclosure gap in devices authorized in 2024 (15.5%).

Reconciliation with our rates. Our race-reporting rate (21.9%) exceeds the 3.6% reported by Muralidharan et al. [5]. The difference arises because the ERA counts any race or ethnicity statement, including qualitative “diverse population” language. This aligns our rate with the 23.7% any-demographics rate of Mehta et al. [6] and, for the most recent authorizations, with the 15.5% of Almarie et al. [10]. The stricter and more comparable figure is the item-2 stratification rate (5.6% for race). The consistency of the transparency picture across five independent audits using different instruments is itself informative: it is the alignment question on which these audits are silent.

The primary result stated without the Crosswalk. 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. Put the other way, the primary finding could be stated without the Crosswalk at all, as “roughly three-quarters of authorized AI devices do not report race,” and the Crosswalk’s contribution is not to generate the proportion but to establish that the devices concerned are aimed at conditions where the omission carries the greatest consequence. That contribution is a matter of which conditions carry documented disparities, which is an epidemiologic fact independent of this study, not a scoring judgement.

Qualifications on the interrupted time-series. The segmented model is fitted over ten annual periods, so it has few observations either side of each knot and the later two knots are effectively unevaluable; the 2025 mean falls below the fitted trajectory, and we would not read a segmented fit this sparse as establishing more than the direction and approximate timing of the 2021 change. The outcome is the mixed-tier composite, so the trend tracks documentation depth broadly and not any single reliable construct. And the December 2024 PCCP guidance was reissued in August 2025 [11], within our window; we did not model the reissue as a separate breakpoint, so the post-2024 segment reflects the PCCP regime as a whole, not one dated intervention.

Supplementary Tables (S1–S10)

FDA specialty panel (lead)	Devices	%
Radiology	1,094	76.5
Cardiovascular	136	9.5
Neurology	65	4.5
Anesthesiology	26	1.8
Gastroenterology–Urology	24	1.7
Hematology	20	1.4
Ophthalmic	10	0.7
Pathology	9	0.6
Clinical Chemistry	9	0.6
Other panels (8)	37	2.6

Table S1: **Cohort by FDA specialty panel** (lead panel; 17 panels). Percentages are calculated over all $N = 1,430$ listed devices, not the 1,428 with retrievable decision summaries. Radiology dominates at 76.5%.

Item	Name	n	%agr.	κ	PABAK	$\kappa_{\text{LLM},R1}$	$\kappa_{\text{LLM},R2}$	Tier
D1.1	Age	153	49.7	0.55	-0.01	0.58	0.73	B
D1.2	Sex	128	73.4	0.73	0.47	0.76	0.81	A
D1.3	Race/ethnicity	153	79.7	0.64	0.60	0.84	0.73	A
D1.4	Disease severity	153	22.2	0.13	-0.56	0.22	0.62	B
D1.5	Geographic site	153	68.6	0.65	0.37	0.68	0.80	A
D1.6	Socioeconomic status	153	92.2	0.32 [‡]	0.84	0.23	0.66	A [‡]
D2.1	Burden asymmetry*	153	54.9	0.07	0.10	n/a	n/a	—*
D2.2	Representativeness*	108	38.0	0.12	-0.24	n/a	n/a	B
D2.3	Geographic diversity	153	69.9	0.71	0.40	0.65	0.58	A
D2.4	Setting match	153	41.8	0.02	-0.16	-0.02	0.34	B
D3.1	Subgroup comparison	153	62.7	0.32	0.26	0.28	0.51	B
D3.2	Limitations	152	25.7	0.01	-0.49	0.01	0.46	B
D3.3	Exclusions	153	56.2	0.07	0.12	-0.07	0.63	B
D3.4	Post-market monitoring	152	98.7	0.87	0.97	0.94	0.81	A
D3.5	PCCP revalidation	1	n/a	n/a	n/a	n/a	n/a	—
D3.6	Adaptive drift monitoring	0	n/a	n/a	n/a	n/a	n/a	—
D4.1	Publication referenced	153	88.2	0.62	0.77	0.42	0.56	A
D4.2	External validation	152	97.4	0.32 [‡]	0.95	0.66	0.49	A [‡]
D4.3	Subgroup analysis [§]	152	100	n.e.	1.00	n/a	n/a	A

Table S2: **Per-item inter-rater and model–human reliability by tier, $n = 153$ validation set.** %agr., percent exact agreement; κ , Cohen’s kappa between the two human reviewers (quadratic-weighted for ordinal items); PABAK, prevalence-and-bias-adjusted kappa; $\kappa_{\text{LLM},R1/R2}$, model versus each human reviewer. **Tier assignment** follows a single criterion, fixed before any hypothesis was tested and applied uniformly to every item: Tier A requires human–human $\kappa \geq 0.60$, or PABAK ≥ 0.75 where the kappa paradox applies. Nine items reach Tier A, constitute the validated core (ERA-9), and carry all inferential claims in the main text; seven do not and are reported descriptively; three items (D2.1, D3.5, D3.6) are not tiered, for the reasons given below. n is the number of validation devices scored by both reviewers with the item applicable; for D2.2, $n = 108$ is the number of devices both reviewers treated as assessable under their own applicability codes. [‡]Kappa-paradox items: near-ceiling agreement with a rare positive class deflates κ ; PABAK confirms high reliability (D1.6, D4.2). [§]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. *In the cohort analysis, D2.1 and D2.2 are computed post hoc from the Epidemiologic Crosswalk and the D2.2 rubric, not from the human quick-codes shown here. D2.1 is therefore not tiered at all: the κ of 0.07 in this row belongs to a human quick-code that the analysis does not use, and D2.1 appears in Table 1 of the main text under a separate Crosswalk-derived marker. Its dependability rests on the empirical near-invariance of the primary outcome to this classification, not on an inter-rater statistic. D3.5 and D3.6 could not be tiered: only one validation device was scored on D3.5 by both reviewers and none on D3.6, both items being not-applicable for the great majority of the cohort. ICC(2,1) for the mixed-tier ERA total = 0.43, which is below accepted thresholds [12] and is the diagnostic that motivated item-level tiering rather than composite reporting. The reliability split is interpretable: items asking whether a named quantity appears in the document were reproduced, while items requiring a coder to judge whether prose constitutes a comparison, a limitation, an exclusion, or an appropriate care setting were not.

Disparity type (D2.1_type)	Conditions	Devices
none (no documented disparity / general-purpose tool)	307	414
mixed (two or more mechanisms)	365	753
access	45	58
detection-gap	10	52
incidence	20	22
reference-standard	11	13
mortality	1	1
Total mapped	759 [†]	1,313

Table S3: **Epidemiologic Crosswalk disparity-type distribution**, at the condition level and device-weighted. [†]Includes 13 human-seeded “known” conditions and 746 model-classified novel conditions. The 115 devices whose target could not be mapped to a specific disease were general-purpose tools (Domain 2 not applicable).

D2.2 subset	<i>n</i>	Score 1	Score 2	Score 3–4	≤ 2 (%)
All assessable	210	50	128	32	84.8
High+medium confidence	108	41	37	30	72.2
High confidence only	35	29	3	3	91.4

Table S4: **D2.2 confidence-stratified sensitivity**. The conclusion that assessable devices under-represent the highest-burden group ($D2.2 \leq 2$) is robust to restricting to higher-confidence model scores. Dichotomization sensitivity: $D2.2 \leq 1$ (severe) = 23.8%; $D2.2 \leq 2$ (moderate or worse) = 84.8%.

	Age	Sex	Race/eth.	Socioecon.	Subgroup analysis
FDA decision summaries (cohort, $n = 1,428$)	56.2	39.7 [‡]	21.9	0.2	15.7
Retrieved primary literature ($n = 57$) [†]	64.9	57.9	28.1	3.5	19.3

Table S5: **Demographic reporting (%): FDA decision summaries versus the retrieved primary literature**. [†]Dual-source subset (68 devices; publications located for 57). Race and socioeconomic status remain sparsely reported even in the peer-reviewed evidence, indicating the gap is not merely a documentation-format artifact. “Subgroup analysis” is D3.1 for the cohort and a reported demographic subgroup performance analysis for the primary literature. [‡]Sex carries a conditional not-applicable rule, so the cohort rate is computed over the 1,321 devices with a sex-applicable indication; the other cohort columns apply to all 1,428.

Item	Name	Devices upgraded	% of 57
D4.1	Peer-reviewed publication exists	34	60
D1.5	Geographic site	15	26
D1.4	Disease severity	11	19
D4.3	Publication reports subgroup performance	10	18
D1.2	Sex	9	16
D4.2	Publication reports external validation	9	16
D1.1	Age	7	12
D1.3	Race/ethnicity	5	9
D2.3	Geographic diversity	5	9
D1.6	Socioeconomic status	1	2

Table S6: **Item-level upgrades from dual-sourcing (Arm 1 $\rightarrow$ Arm 2), 57 located devices.** Number of devices whose consensus code increased when the retrieved primary publication was added to the source set. The gains concentrate in the existence of a publication (D4.1) and in reporting axes orthogonal to population-data alignment; the two items most relevant to representativeness, race (D1.3) and socioeconomic status (D1.6), barely move (bold). All 5 race upgrades occurred in devices targeting a high-disparity condition that was previously unassessable, rescuing 5 of the 27 such devices for representativeness assessment. Domain-3 device/design items, D2.1, the graded D2.2 value, and D2.4 (setting) were carried from Arm 1 unchanged and are omitted.

Pathway	LLM cohort median ERA %	Human-consensus median ERA % ($n = 153$)
510(k)	12.5	14.6
De Novo	39.3	36.4
PMA	36.4	32.6

Table S7: **Cross-validation of cohort-wide extraction against the human-adjudicated subset.** Model-derived per-pathway medians closely track the human-consensus medians, supporting cohort-wide application. Additional Domain 2 sub-item frequencies: geographic diversity (D2.3) scored 0/1/2 in 60.2/39.7/0.1% of devices; setting match (D2.4) in 71.4/28.6/0.0%. No device described validation in community, safety-net, or LMIC settings appropriate to the condition’s care pathway.

Target condition	Devices	Highest-burden group(s)	Approximate U.S. burden
Intracranial haemorrhage / stroke	61	Black, Hispanic	~795,000 strokes/year (~8–15% haemorrhagic); first-stroke risk ~2× higher in Black adults
Coronary artery disease	38	South Asian, Black; low SES	~16.3 million adults with coronary heart disease; earlier onset and higher mortality in Black and South Asian adults
Lung cancer / lung nodules	36	Black men; low SES	~235,000 new cases/year; historically highest incidence in Black men; screening under-use in low-SES
Atrial fibrillation	27	Black (under-diagnosed)	~10.5 million adults; under-diagnosed in Black patients despite higher stroke risk
Breast cancer (screening)	19	Black women	~311,000 new cases/year in women; ~40% higher mortality in Black women
Cardiac arrhythmias	14	Black adults	Overlaps atrial fibrillation (~10.5 million); higher sudden-cardiac-death rates in Black adults
Breast cancer (diagnosis)	14	Black women (mortality)	See breast cancer (screening)
Prostate cancer	13	Black men	~299,000 new cases/year; ~70% higher incidence and ~2× higher mortality in Black men
Incidental pulmonary embolism	12	Black adults	Higher venous-thromboembolism incidence and case-fatality in Black adults
Pulmonary nodules	11	Black adults	Lung-cancer incidence/mortality disparity (see above)
Pneumothorax	10	Men; low-SES adults	Higher incidence in men; access gaps in low-SES adults
Colorectal polyps / cancer	10	Black adults	~153,000 new colorectal cancers/year; earlier onset and ~40% higher mortality in Black adults
Dental caries	10	Low-SES, Hispanic, Black children	Most common chronic disease of childhood; concentrated in low-SES and minority children
Diabetic retinopathy	8	Black, Hispanic, Native American; low SES	~9.6 million adults with diabetic retinopathy (2021); 2–3× higher in affected groups

Table S8: Frequently targeted high-disparity conditions, with U.S. disease burden. The 13 most common high-disparity conditions by device count, together with diabetic retinopathy (8 devices), which is carried as a running example in the main text and ranks just below the top group. Together they account for 283 of the 899 high-disparity-condition (D2.1=1) device authorizations (31%). Burden figures are approximate, rounded U.S. estimates and are illustrative of scale, not exhaustive; disparity direction and highest-burden group follow the Epidemiologic Crosswalk classification. Sources: cancer incidence from American Cancer Society *Cancer Statistics, 2024* (Siegel et al.) and cancer disparities from *Cancer Statistics for African American and Black People, 2025* (Saka et al.); stroke and coronary heart disease from the American Heart Association *2024 Heart Disease and Stroke Statistics* update; atrial fibrillation from the 2024 *J. Am. Coll. Cardiol.* minimum-national-prevalence estimate (≥ 10.55 million); diabetic retinopathy from the CDC/IHME *JAMA Ophthalmology* 2023 estimate; dental caries from CDC NCHS. Prevalence and incidence figures span different measures and should not be summed. The devices for which representativeness cannot be assessed (main text) target conditions that collectively affect tens of millions of U.S. adults and carry substantially higher incidence or mortality (roughly 1.4- to 3-fold, depending on the condition) in Black, Hispanic, or Native American populations.

Panel A. Known-groups contrast: decision-summary ERA% by independent literature label			
Known group (from primary literature)	n	Human ERA% median (range)	Model ERA% median (range)
Equity-attentive (paper reports race <i>and</i> subgroup)	8	50.9 (16.7–73.1)	50.0 (16.7–64.3)
Neither reported	38	25.0 (0.0–66.7)	28.6 (4.2–64.3)
Mann–Whitney U , equity-attentive > neither		$p = 0.003$	$p = 0.007$
Panel B. Convergent validity: literature equity-evidence count (0–4) vs ERA%			
Spearman ρ (all located devices)	n	Human score	Model score
Correlation with decision-summary ERA%	56–57	$\rho = 0.49, p < 0.001$	$\rho = 0.47, p < 0.001$
Panel C. Item-level concordance: race reporting across the two documents			
	Paper reports race	Paper does not	
FDA summary reports race (D1.3 $\geq$ 1)	11	12	
FDA summary does not	5	29	
Fisher's exact (concordance)		OR = 5.3, $p = 0.008$	

Table S9: **Criterion and convergent validity of the ERA against the independent primary literature** ($n = 57$ dual-source devices with a located publication). The ground-truth equity label is derived from the retrieved pivotal publication; the ERA score under test is derived from the FDA decision summary only. Because the label and the score come from separate documents, the associations are not mechanically induced. The equity-attentive group is small ($n = 8$), so the continuous correlation in Panel B is the more robust estimate. Eleven devices reporting one of the two criteria (race or subgroup, not both) are omitted from the Panel A dichotomous contrast. Computed by `known_groups_validity.py` (deposited).

ID	Protocol specified	What was done	Bearing on primary outcome
D1	De Novo cases coded from the full Decision Summary	26 of 37 initially retrieved as the short Classification Order; re-downloaded and re-extracted (Note 1)	None. Corrected before analysis; human coding used Decision Summaries throughout
D2	Complete Crosswalk built and deposited before D2 scoring; disparity type assigned by two human reviewers	13 anchor conditions human-classified; 746 novel conditions classified by the model under the deposited expansion procedure (Note 3)	Bounded. Defines the stratum only; primary outcome near-invariant (76.6% vs 78.1% cohort-wide)
D3	D2.2 scored by dual human review	D2.2 for 210 assessable devices scored by the model against the Part E rubric (Note 4)	None. Tier-B item; excluded from inference
D4	Sampling parameters: temperature not set	Temperature set to 0 (Note 2)	None. Affects reproducibility description only
D5	Triangulation sub-sample $n = 50$; second source = labeling/IFU	Dual-source arm $n = 68$ (57 located); second source = peer-reviewed pivotal publication (Note 6)	None on the primary outcome. Changes what the secondary test demonstrates; per-pathway labeling concordance not reported
D6	Primary outcome = D2.2 distribution among D2.1=1 devices	Primary outcome = proportion of that stratum unassessable because race unreported; original reported unchanged	Definitional. See note above
D7	Any item failing the κ threshold triggers expansion of human review to 40% of cohort	Not executed; not triggered for the primary-outcome item; capacity unavailable for Tier-B items	None on the primary outcome, which passed the rule as specified
D8	No item-level reliability tiering specified	Two-tier scheme introduced post hoc; inference restricted to Tier A; all items reported	Restricts claims; does not alter any computed value
D9	No known-groups validity analysis specified	Post-hoc criterion-validity analysis against the primary literature added (Table S9)	None. Reported as supporting construct validity, not as a test

Table S10: **Complete register of deviations from the deposited pre-registration.** Data-quality corrections made before analysis (40 out-of-range binary entries; 43 applicability-disagreement cells) are described in Note 5 and logged separately, as they correct coder entry and are not departures from protocol. No deviation alters the primary outcome’s computed value; D6 alters which quantity is designated primary, and D7 and D8 concern remedies for unreliable items rather than the reliable item on which the primary outcome rests.

Supplementary 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] 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>.
- [3] 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/>.
- [4] Anthropic. Model IDs and versioning. Claude Platform documentation, 2026. URL <https://platform.claude.com/docs/en/about-claude/models/model-ids-and-versions>.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.

- [11] 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-intel>
- [12] 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.