Research Square

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

Public regulatory data infrastructure for lifecycle oversight of AI medical devices

Yu Han

yu.han@eng.ox.ac.uk

University of Oxford

Aaron Ceross

University of Birmingham

Article

Keywords: artificial intelligence, medical devices, software as a medical device, regulatory databases, post-market surveillance, lifecycle governance, public data infrastructure, health technology assessment

Posted Date: June 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10095369/v1

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

Additional Declarations: No competing interests reported.

Public regulatory data infrastructure for lifecycle oversight of AI medical devices

Yu Han[1,*], Aaron Ceross[2]

[1]*Department of Engineering Science, University of Oxford, Oxford, UK*

[2]*Birmingham Law School, University of Birmingham, Birmingham, UK*

[*]*Corresponding author: yu.han@eng.ox.ac.uk*

Abstract

Regulatory science for AI medical devices has moved toward lifecycle oversight: predetermined change-control plans, living device databases, and machine-learning standards assume software evolves after market entry. The public records through which hospitals, safety officers, and researchers identify devices, trace change, read evidence, and link authorisation to safety signals have not kept pace, remaining built to confirm market authorisation. We convert previously described registry limitations into a source-grounded public-data specification (five public-governability functions and 20 structural requirements) and audit public records in the USA, China, the EU, Japan, and Australia. Only 6% reached structured, queryable availability; deficits concentrated in software change, evidence, linkage, and AI metadata, and held under a conservative 12-item core. The most distinctive deficit: AI-specific attributes (model class, training-data modality, provenance) are not integrated as structured fields. Lifecycle governance requires public records encoding device identity, change, evidence, linkage, and AI attributes in structured form.

Keywords: artificial intelligence, medical devices, software as a medical device, regulatory databases, post-market surveillance, lifecycle governance, public data infrastructure, health technology assessment

1 Introduction

A hospital evaluating an AI-enabled radiology triage system can confirm from public regulatory records that the device was authorised, but typically not which software version is currently marketed, whether the model has changed since authorisation, what population it was validated on, or whether its authorisation record can be linked to recalls or adverse-event reports. These are not questions about internal regulatory review; they are questions about whether public regulatory data can support safe procurement, deployment, monitoring, and accountability after a device reaches the clinic. AI deployment in healthcare depends on more than the authorisation itself [Topol, 2019, Rajpurkar et al., 2022, Wiens et al., 2019]: procurement committees compare candidate devices, safety officers update local clinical safety cases when manufacturers change software [Char et al., 2018, Sendak et al., 2020], and post-market surveillance must connect authorisation records to recalls and adverse-event signals. AI systems also raise distribution-shift and portability concerns that registration-oriented registries were never designed to expose: a model trained on one population may underperform in another [Finlayson et al., 2021, Roberts et al., 2021], validation evidence often fails to generalise across hospitals [Seyyed-Kalantari et al., 2021, Obermeyer et al., 2019], and a model that adapts after deployment may diverge from the version originally evaluated [Babic et al., 2019, Vokinger et al., 2021]. Public registry architecture is the baseline infrastructure for external accountability around these concerns [Hwang et al., 2019, Gilbert et al., 2023]. The argument is not that public regulatory records should replace confidential regulatory review, vendor documentation, local clinical assurance, or health technology assessment; it is that all of these activities depend on a shared public reference layer: a stable device identity, a current version, a traceable authorisation record, an accessible evidence context, and a route from authorisation to post-market signals. Without that layer, hospitals and safety actors can verify that an AI device entered the market but cannot reliably determine from public records what version is in use, what evidence supports it, how it has changed, or whether later safety signals attach to the same regulated object.

Major regulators have moved their oversight philosophy toward the lifecycle. The FDA's final guidance on predetermined change control plans (PCCPs), published in August 2025, accepts that AI-enabled device software will change after deployment and establishes a procedure for pre-authorising categories of modification [FDA, 2025]; by early 2026, 15 AI/ML-enabled devices had been authorised with PCCPs [FDA, 2025]. The European Commission describes EUDAMED as a "living picture of the lifecycle" of medical devices [European Commission, 2025], and the IMDRF has issued terminology standards for machine-learning-enabled devices that presuppose lifecycle-aware data management [IMDRF, 2022]. The shared premise is that AI and software devices are not static hardware objects.

The public-facing registries through which these systems expose device records have not moved in the same direction. The FDA's 510(k) database, the NMPA's UDI system, the ARTG, and the PMDA's approval database remain primarily *registration-oriented*: they are built to answer whether a device is legally marketed, and their data models centre on device name, manufacturer, classification, and approval status rather than on the structured lifecycle fields needed to trace mutable software after authorisation. That design was adequate when the regulated object had a stable identity, a performance fixed by its physical design, and post-market behaviour that could be assumed to reflect pre-market evaluation. None of these assumptions hold for an imaging algorithm that is updated between submissions, whose performance depends on its training distribution, and whose deployed behaviour can drift as patient populations shift.

Recent work has documented the resulting problems. Ong et al. (2026) identify four challenges in AI-as-a-medical-device databases: inconsistent identification, limited version tracking, insufficient evidence transparency, and weak cross-database linkage [Ong et al., 2026]; the FDA's own AI-Enabled Device List states that it is "not comprehensive" and is built primarily from AI-related terms in public summaries and

product classifications [FDA, 2025]; and AI devices have been shown to be hard to identify systematically across the US, EU, and China [Liu et al., 2024, Muehlematter et al., 2021]. These findings establish that current registries are inadequate, but they describe the problem as one of data quality or access (missing values to fill, doors to open) rather than asking the constructive question they imply: *what would an adequate registry contain?* We treat this as a design problem. Existing architectures were built to establish legal market status; as implemented, they do not expose the structured lifecycle fields needed to identify, compare, trace, and monitor mutable AI software through public records.

The contribution is therefore not the observation that AI medical-device registries are incomplete; that problem is now recognised. The contribution is to convert that problem into an auditable public-data specification: five governance functions, 20 structural requirements, a conservative 12-item core, a gap typology that distinguishes missing fields from unpublished fields, unstructured disclosures, linkage failures, and population failures, and a candidate data extension that maps each proposed field to an audited governance requirement. Relative to Ong et al.'s account of recurring registry challenges [Ong et al., 2026], the present study moves from problem description to a source-grounded specification that can be applied reproducibly across jurisdictions and read directly as a reform target.

This paper therefore asks what a public regulatory record environment must structurally contain for software and AI devices to be identifiable, comparable, traceable through change, and connectable to post-market signals by users beyond the regulator. We answer in three steps. First, drawing on public-administration scholarship on transparency and information infrastructure, regulatory guidance, and prior empirical work, we derive a source-grounded specification of five governance functions and 20 structural requirements, justified through a four-anchor source-concordance traceability matrix (regulatory, governance, clinical/operational, and empirical/problem anchors) that substitutes transparent derivation for an expert-validation procedure we did not run (Section 2). Second, two annotators independently audit the public regulatory record environments of five jurisdictions against the specification, scoring each cell on a 0–3 availability scale, classifying unmet cells into five gap types, and assigning per-cell confidence; the audit is stress-tested under two sensitivity restrictions (core-only, High/Medium-confidence) and three pre-specified empirical probes (FDA AI identifiability, FDA product-family traceability, China findability) (Sections 3 and 5). Third, we propose a candidate minimum public data extension aligned to the audited requirements and tier reform recommendations by gap type and confidence (Section 4). We do not claim the specification has been externally validated or that it represents expert consensus; we claim that its derivation is auditable, its application reproducible from the deposited materials, and the patterns it surfaces stable under the conservative core specification and the high-confidence subset.

2 Conceptual framework

2.1 The insufficiency of "more transparency"

It is tempting to frame registry inadequacy as a transparency problem and to propose "more transparency" as the solution. But transparency is not a single property; it is a family of design choices about what information is made visible, to whom, in what form, and for what purpose [Hood and Heald, 2006, Heald, 2006].

Hood and Heald's (2006) typology distinguishes transparency of rules, processes, outputs, and outcomes. A device registry that lists approved devices provides *output transparency* (the public can see *that* a device was approved), but output transparency alone does not make a device *governable*: it does not equip public users to identify, compare, evaluate, and monitor the device through its lifecycle.

For static hardware this may suffice: knowing that a hip implant was approved, by whom, and in what class tells a clinician most of what they need. For AI-enabled software it does not. A user who can see that

"Algorithm X was cleared via 510(k) in 2024" but cannot determine its current version, training data, modification history, or post-market performance lacks the information to make an informed adoption, purchasing, or surveillance decision.

The public users we have in mind are concrete: a hospital procurement committee comparing candidate devices, a clinical safety officer updating a local safety case after a software change, a health technology assessment body or researcher comparing evidence and performance claims, a patient-safety group linking authorisation records to recalls and adverse-event reports, and a regulator in another jurisdiction comparing device evidence. We do not claim these actors currently treat public regulatory records as their primary source; procurement and HTA also draw on vendor submissions, confidential documentation, and direct engagement. The claim is infrastructural: if public records are to support external comparison, monitoring, and accountability, they must expose the relevant identity, evidence, change, and linkage fields in structured form.

We therefore propose that the appropriate design target for AI device public regulatory record environments is not transparency *per se* but **public governability**. In plain terms, public governability is a practical property of a public data system: whether an external user can answer the minimum questions needed to procure, deploy, monitor, and investigate an AI medical device. More formally, it is the structural capacity of a public information system to support the epistemic actions (identification, comparison, tracing, linking, and characterisation) that governance of the regulated object requires. This builds on two further threads in public-administration scholarship. Meijer [Meijer, 2013] argues that transparency is constituted in the relationship between an information system and the actors who must use it, not in the act of disclosure alone; what matters is whether disclosed information is usable for the governance task at hand. Bowker and Star [Bowker and Star, 1999, Star and Ruhleder, 1996] show that what a registry can express is bounded by the categorical vocabulary of its data model: information that has no field cannot be governed through that system. Public governability is the operational consequence: a system is publicly governable to the extent that it discloses, in usable structured form, the categories of information the relevant governance task requires.

2.2 Five governance functions

We derive five functions that a public regulatory record environment must support if software and AI devices are to be governable. Each function corresponds to a specific epistemic need and maps to a recognised variety of public transparency. These functions are not arbitrary: they follow from what public oversight of mutable, software-intensive regulated objects structurally requires, and from what regulatory systems themselves claim to provide.

2.2.1 F1 Identity: what is the regulated object?

Transparency variety: output transparency (what was approved)

The most basic governance function is identification: can a public user determine what the regulated object actually is? For AI devices, this requires distinguishing software from hardware, standalone software (SaMD) from software embedded in hardware (SiMD), and AI-enabled software from rule-based or deterministic software. It also requires that the device's intended clinical function (what it does for the patient) be recorded in a form that a non-specialist can understand and a researcher can query.

The IMDRF's SaMD classification framework presupposes this distinction [IMDRF, 2013], and regulatory bodies across the five jurisdictions audited here have adopted SaMD as a regulatory category. The audit reports below to what extent this regulatory concept has been translated into a structured field within each jurisdiction's public regulatory record environment.

2.2.2 F2 Version and change legibility: has the object changed, and under what authority?

Transparency variety: process transparency (how the regulated object evolved)

Software change is constitutive, not incidental: updateability is a defining feature of software-based devices and the explicit basis for the FDA's PCCP framework [FDA, 2025]. A registry that records only the initial authorisation presents a *snapshot* as if it were the *current state*. This is acute under a PCCP, which pre-authorises categories of change so that planned modifications to an AI device's algorithm, training data, or intended use may occur without a new submission, and thus without a new public record. Unless the registry independently tracks versions and their regulatory basis, a public user cannot know whether the deployed device still resembles the one evaluated.

2.2.3 F3 Evidence legibility: what was the basis for regulatory confidence?

Transparency variety: reason transparency (why this decision was made)

Public governance requires not just that decisions are visible but that their *reasons* are recoverable [Hood and Heald, 2006]: the intended use evaluated, the clinical setting and population in which the device was validated, and the performance metrics (sensitivity, specificity, AUC, or equivalent) achieved. Without these, a public user cannot judge whether a device validated in one population and setting transfers to another, a researcher cannot compare competing devices' evidence bases, and post-market surveillance cannot tell whether real-world performance has diverged from the validation benchmark.

2.2.4 F4 Linkage: can pre-market records be connected to post-market reality?

Transparency variety: outcome transparency (what happened after approval)

Lifecycle regulation treats pre-market evaluation as the beginning, not the end, of oversight: post-market surveillance (adverse events, recalls, field safety corrective actions) generates the outcome data that ongoing governance depends on. But this data sits in separate systems (FDA MAUDE, TGA DAEN, the EU vigilance module). A registry supports governance through linkage when a public user can connect an authorisation record to its post-market events, which requires a persistent device identifier (the UDI-DI is designed for exactly this) used consistently across pre- and post-market databases.

2.2.5 F5 AI specificity: can you tell what kind of AI this is?

Transparency variety: object transparency (what is this thing made of)

F5 asks whether the public record exposes AI-relevant technical metadata (model/method class, AI task category, training/validation data modality and provenance) as structured registry fields rather than as supplementary catalogues separate from the primary authorisation records. Three items are essential and each has a concrete governance use: *algorithm type* (deep learning vs. classical machine learning vs. rule-based logic), because models differ in their susceptibility to distribution shift; *AI task* (diagnosis, triage, image enhancement, risk prediction, treatment planning), needed for comparison and health technology assessment; and *training-data domain* (modality, population, geographic context), needed to assess portability of performance claims.

This is not merely a gap but a **field-level uptake failure**: the vocabulary is not missing (IMDRF and related standards already define terminology for machine-learning-enabled medical devices), but these concepts have not been translated into structured, queryable fields in primary public records. The FDA AI-Enabled Device List and the TGA AI list are the closest approximations, but both are supplementary catalogues, not fields within the primary records, and neither exposes model/method class or training-data provenance as queryable data. Such records can show that a device exists and was approved, yet

cannot express whether it uses a convolutional neural network or a random forest, was trained on 500 or 500,000 images, or adapts after deployment.

2.3 From functions to requirements

Each governance function is operationalised through four structural requirements, yielding 20 requirements in total (Table 1). Each cell (jurisdiction × requirement) is scored on a four-point ordinal availability scale, classified by gap type when below the structured threshold, and assigned a per-cell confidence rating. The full scoring rubric and gap-type definitions are given in the Methods (Section 5); we summarise here. A score of 0 indicates that the information is not publicly available; 1, that it is public only in narrative, PDF or free-text form; 2, that it is a structured field but not queryable / downloadable / consistently accessible; 3, that it is a structured, queryable or downloadable field. This four-point scale is preferred over a three-point scale because it separates "public somewhere" (score 1) from "structured but not yet usable as data" (score 2), a distinction that the policy implications turn on.

The 20 requirements are organised as a 12-item conservative *core* specification, the *full* 20-item set, and an unscored *extended* discussion set; the audit reports findings under both the full and core specifications, with the core serving as the conservative-sensitivity restriction. The justification for inclusion of each requirement is recorded in a four-anchor source-concordance traceability matrix (regulatory, governance, clinical/operational, empirical/problem anchors), provided in the supplementary material.

The four items at F4d, F5b, F5c and F5d are anchored by governance, clinical/operational, and empirical/problem considerations but not by an existing regulatory submission requirement that names them as a structured field; they are retained in the full specification with explicit Medium inclusion confidence and discussed accordingly throughout the analysis.

2.4 Gap typology: five categories distinguished by reform target

Not all gaps in public regulatory records imply the same underlying problem, and the nature of the gap determines the appropriate policy response. We therefore developed a five-category gap typology that refines the schema/access/population framework used in earlier studies. The refinement was motivated by the observation that conceptually distinct deficiencies were being grouped together despite requiring fundamentally different remedies.

A schema gap occurs when no public field exists and there is no evidence that the regulator collects the information in structured form. Addressing such gaps requires the introduction of new data elements, controlled vocabularies, and potentially revised submission requirements. A publication–access gap arises when guidance documents or submission specifications indicate that information is collected or reviewed, but the corresponding data are not exposed through public regulatory records. In these cases, improved transparency rather than additional data collection is required. A structuring gap describes information that is publicly disclosed but only in narrative text, PDFs, or other unstructured formats, limiting its usability as data despite its availability. A linkage gap reflects the inability to reliably connect related records across systems because of the absence of stable identifiers. Finally, a population gap occurs when structured fields exist but are incompletely or inconsistently populated, pointing to shortcomings in data-quality enforcement.

Distinguishing between these gaps has important implications for regulatory reform. Information buried in PDFs requires different interventions from information that regulators collect but do not publicly disclose, even though both appear as unavailable from the perspective of downstream users. Likewise, absent cross-system identifiers and poorly populated fields represent different failures of public data infrastructure despite producing similar practical barriers to analysis. Conflating these scenarios risks mischaracterising both the scale and the nature of the problem.

Table 1: Proposed public-governability audit instrument: five governance functions and 20 structural requirements (full specification). The framework is a candidate minimum public-data specification derived through transparent source concordance, not an externally validated consensus standard. The "Anchor type" column summarises which of the four source-concordance anchors (R = regulatory, G = governance, C = clinical/operational, E = empirical/problem) support inclusion. The "Core" column flags the 12 items in the conservative core specification used in sensitivity analysis; the four non-core items lacking a direct regulatory-submission anchor (F4d, F5b, F5c, F5d) are retained with Medium inclusion confidence. The full traceability matrix is in the supplementary material.

Function	**ID**	**Requirement**	**Anchors**	**Core**
F1: Identity	F1a	Software / AI status	R G C E	✓
	F1b	Stable public identifier	R G C E	✓
	F1c	Intended use / clinical indication	R G C E	✓
	F1d	Clinical function / use mode	R G C E	
F2: Version & Change	F2a	Current software / model version	R G C E	✓
	F2b	Public change / update history	R G C E	✓
	F2c	Regulatory basis for change	R G C E	✓
	F2d	Locked / periodically updated / adaptive / continuous-learning status	R G C E	
F3: Evidence Legibility	F3a	Public decision / evidence summary	R G C E	✓
	F3b	Validation setting	R G C E	✓
	F3c	Validation population and data modality	R G C E	
	F3d	Performance metrics and stated limitations	R G C E	✓
F4: Linkage	F4a	Current certificate / authorisation status	R G C E	
	F4b	Recall / safety notice linkage	R G C E	✓
	F4c	Adverse-event / incident linkage	R G C E	✓
	F4d	Product-family / predecessor–successor linkage	G C E	
F5: AI Specificity	F5a	AI task category	R G C E	✓
	F5b	Model / method class	G C E	
	F5c	Training / validation data modality	G C E	
	F5d	Training / validation data provenance summary	G C E	

Importantly, strong claims that regulators "already hold the data" are justified only for confirmed publication–access gaps. They do not apply to schema gaps, which require new data collection, or to structuring gaps, which require improved data representation. We therefore align our reform recommendations with the specific gap types identified, recognising that different deficiencies demand different policy responses. Table 2 summarises the five gap types, the diagnostic question that distinguishes each, the matching reform instrument, and the strength of claim the audit supports for each.

Table 2: The five-way gap typology: each gap type is distinguished by a diagnostic question and maps to a different reform instrument. The typology is reusable beyond AI medical devices, wherever public records are audited for governance capacity.

Gap type	**Diagnostic question**	**Reform instrument**	**Claim strength**
Schema	Does any public field exist, and is there no evidence the regulator collects it in structured form?	New data element + controlled vocabulary; revised submission requirement	Strong on absence; cautious on "regulator does not hold"
Publication–access	Is the information collected/reviewed per guidance but not publicly exposed?	Transparency mandate (publish already-held data)	Strong only for High/Medium-confidence cells
Structuring	Is it public but only as narrative/PDF/free text?	Structure already-disclosed content into queryable fields	Strong (disclosure observed directly)
Linkage	Can related records be connected across systems via a stable identifier?	Identifier infrastructure / cross-registry keys	Strong where identifier demonstrably absent
Population	Do structured fields exist but sit empty or inconsistent?	Data-quality enforcement on existing fields	Strong where field exists and is unpopulated

3 Results

Across the full 100-cell audit, the public regulatory record environments we examined supported registration but little of what lifecycle governance requires. Ninety-four of 100 cells fell below the structured, queryable threshold, and the deficits were not evenly distributed: they concentrated in software version and change (F2), evidence legibility (F3), and, most severely, AI specificity (F5). Within primary public records, no jurisdiction exposed model or method class, training-data modality, or data provenance as a structured field; every F5b–d cell scored 0 or 1 (absent or narrative only). Four of the six cells reaching the top of the availability scale are basic device-identity fields (F1). We report the audit at function level rather than as a jurisdictional league table, because the requirement- and function-level deficit patterns, not the country rankings, are the substantive finding; jurisdictional totals are deposited as diagnostic material (Supplementary Appendix D).

Five findings structure the results. First, public availability is not the same as public data infrastructure: many items are visible only through narrative summaries, PDFs, free text, or manual inference. Second, the deficits are functional rather than merely jurisdictional, concentrating in version and change traceability, evidence legibility, post-market linkage, and AI-specific metadata. Third, the USA is closest to structured exposure, but even the richest public API environment does not expose current software version, structured product-family lineage, or training-data provenance. Fourth, most unmet cells are not confirmed publication–access gaps, so the remedy is not simply to publish data already held by regulators. Fifth, the three targeted probes make the architectural finding observable in concrete public

data systems: known FDA AI devices are difficult to identify from primary records, China's bulk UDI infrastructure is AI-blind at the field level, and FDA product-family lineage is not exposed as structured data even in openFDA.

3.1 Source inventory and evidence depth

We audited 24 public regulatory sources across the five jurisdictions, locally archived on 2026-05-01 (HTML capture + screenshot + `meta.json` per source); the source base was locked at this archive for both pilot and full coding (full inventory and per-source capture notes in Supplementary Appendix A). Four dates recur below: the *audit cut-off* (2026-05-01); the *source-archive lock* (local capture as of that date); the *stress-test reference-data download* (2026-05-14, for the FDA AI-list CSV and the NMPA UDI full release); and the *post-review correction* to the China UDI evidence. Sources span three evidence-depth tiers: bulk-downloadable structured data (FDA 510(k), PMA, De Novo, MAUDE and Recalls via the openFDA API; the FDA AI-Enabled Device List CSV; and the NMPA UDI full release, a bulk XML release); search-interface-only structured data (PMDA Shonin, the ARTG, and the four EUDAMED modules becoming mandatory after the cut-off); and documentation-only sources (the EUDAMED Vigilance and Market Surveillance modules, not yet operational at cut-off). Capture-level problems, namely TGA pages behind an Akamai CDN that refuses non-browser agents, NMPA landing pages rescued via the Wayback Machine, and the EUDAMED single-page-application routes that are not separately archivable, are recorded per source in Supplementary Appendix A and do not affect the bulk-data findings, which rest on directly retrieved releases. Evidence-depth propagates into per-cell confidence: directly observed structured data carries High confidence; documentation alone carries Medium; ambiguous documentation, translation uncertainty, Wayback substitution, or interface inconsistency carries Low.

3.2 Inter-annotator reliability

Pre-adjudication agreement on the full 100 cells is reported before any substantive findings. We report raw agreement and linear-weighted Cohen's κ for the ordinal availability score, raw agreement and Cohen's κ for gap type with Gwet's AC1 supplementing where category prevalence in either annotator's coding exceeded 60%, and raw agreement for confidence rating.

Table 3: Inter-annotator reliability on the full 100 cells, pre-adjudication. Interpretation thresholds: $\geq$ 0.80 strong, 0.67–0.79 acceptable, 0.50–0.66 cautious, $<$ 0.50 codebook failure.

Coding dimension	**Raw agreement**	**Cohen's κ**	**Gwet's AC1**	**Interpretation**
Availability score (0–3, linear weights)	0.780	0.724	n/a	Acceptable
Gap type (5 categories)	0.796	0.703	not triggered (<60% prevalence)	Acceptable
Confidence rating (3 categories)	0.740	0.555	n/a	Cautious

The full disagreement pattern (which gap-type pairs are confused, which jurisdictions concentrate disagreements, which boundary rules surfaced) is reported in Supplementary Appendix B; Table 3 reports the headline statistics. The adjudication log is in Appendix C. As an adjudication diagnostic, post-revision Cohen's κ on the availability score is 0.854 (raw 0.890, Strong), reported in the Supplementary Information; the pre-adjudication figure remains the headline measure of independent coder agreement as is standard practice.

3.3 Core specification: conservative-sensitivity result

Before reporting the full audit, we report findings under the 12-item core specification. The core specification is the conservative-sensitivity restriction: a finding that fails to hold under the core does not

survive the most defensible version of the audit.

The structural deficits identified under the full specification (Section 3) are **stable** under the core. In the core 60-cell subset, 52 of 60 cells (87%) scored 0 or 1 and only 5 of 60 (8%) scored 3 (three in the USA: F1a, F1b, F5a; one in Australia: F1b; one in China: F1b). Function-level deficits at F2 (Version & Change), F3 (Evidence) and F5 (AI Specificity) hold under both the full and core subsets. All five core score-3 cells are F1 (Identity) entries, three of them in the USA; the AI-specificity, version-traceability, and evidence-legibility deficits remain asymmetric, with the USA closer to structured exposure than the other four jurisdictions. The cells driving the core findings are dominated by High and Medium confidence ratings (15 of 60 High and 33 of 60 Medium); 12 of 60 core cells are Low confidence, concentrated in Japan (translation-mediated reading of PMDA review reports), China F2–F5 cells (where the 2026-05-01 UDI bulk release upgrades F1b (the stable public identifier) but does not extend to F1a software/AI-status, software-version, validation or training-data fields), and Australia F4c (inaccessible adverse-event source at cut-off).

3.4 Full architecture audit

The five public regulatory record environments differ in structure (Table 4). The unit of comparison is the public record environment available to an external user, not the regulator's internal information system; the audit therefore measures public governability, not regulatory competence. Read this way, the audit compares public-data affordances across jurisdictions rather than ranking legal systems.

Table 4: Public regulatory record environments compared. "AI-specific public list" refers to a dedicated agency-level AI/ML device list; "public evidence summaries" to routinely published decision/evidence documents; evidence-depth tiers follow the Methods (bulk-downloadable / search-interface / documentation-only).

Jurisdiction	**Primary public record environment**	**AI list**	**Public evidence summaries**	**Adverse-event / recall source**	**Evidence-depth tier**
USA (FDA)	510(k) / PMA / De Novo	Yes	Yes (decision summaries)	MAUDE; Recalls (openFDA)	Bulk-downloadable (openFDA)
EU (EUDAMED)	EUDAMED modules	No	Held by Notified Bodies, not published	Vigilance module (not yet mandatory)	Search-interface / documentation-only
China (NMPA)	Registration-certificate DB; UDI	No	Technical review reports (narrative)	Limited public	Bulk (UDI) + search (certificates)
Japan (PMDA)	Shonin approval DB	No	Review reports (narrative, translated)	Limited public	Search-interface
Australia (TGA)	ARTG	Yes	Limited	DAEN	Search-interface (CDN-restricted)

Table 5 reports adjudicated availability scores across the full 20-requirement specification for all five jurisdictions on the 0–3 scale. Aggregate jurisdiction totals are provided in Supplementary Appendix D rather than the main text, on the rationale that requirement-level and function-level deficit patterns are the substantive findings; jurisdictional rankings are diagnostic, not the conclusion.

Figure 1 provides the full score matrix (0–3 availability per cell); per-cell confidence ratings are

Table 5: Adjudicated availability scores under the full 20-requirement specification. Scale: 0 = not publicly available; 1 = public only in narrative / PDF / free-text / inferential form; 2 = structured but not queryable / downloadable / consistently accessible; 3 = structured, queryable or downloadable. Per-cell confidence ratings and gap types accompany each non-3 score in Supplementary Appendix D.

Function	Requirement	USA	EU	China	Japan	Australia
F1: Identity	F1a: Software/AI status	3	1	0	1	2
	F1b: Stable public identifier	3	2	3	2	3
	F1c: Intended use / clinical indication	1	1	1	1	1
	F1d: Clinical function / use mode	1	1	1	1	1
F2: Version & Change	F2a: Current software/model version	1	0	0	0	0
	F2b: Public change/update history	1	0	0	0	0
	F2c: Regulatory basis for change	1	0	0	0	0
	F2d: Locked/adaptive/CL status	0	0	0	0	0
F3: Evidence Legibility	F3a: Public decision/evidence summary	1	1	1	1	0
	F3b: Validation setting	1	1	1	1	0
	F3c: Validation population/data modality	1	1	1	1	0
	F3d: Performance metrics + limitations	1	1	1	1	0
F4: Linkage	F4a: Current certificate/auth status	3	2	2	2	2
	F4b: Recall/safety notice linkage	1	0	1	1	1
	F4c: Adverse-event/incident linkage	1	0	0	0	0
	F4d: Product-family/successor linkage	1	0	0	0	0
F5: AI Specificity	F5a: AI task category	3	0	0	0	1
	F5b: Model/method class	0	0	0	0	0
	F5c: Training/validation data modality	1	0	0	1	0
	F5d: Training/validation data provenance	1	0	0	1	0

reported in Figure 2 and Supplementary Appendix D, not overlaid on Figure 1.

We describe the audit findings at function-level rather than as a jurisdictional ranking. Function-level deficits anchored in High-confidence cells across at least three jurisdictions are reported as the substantive findings of the audit; jurisdictional totals are provided in Supplementary Appendix D as diagnostic, not as a league-table conclusion.

To avoid resting the headline on a single threshold, we report availability at three cut-points. Information was public *in any form* (score $\geq$1) in 55 of 100 cells; *structured in some form* (score $\geq$2) in 13 of 100 cells; and *machine-actionable*, structured and queryable or downloadable (score 3), in only 6 of 100 cells. The direction of the finding is the same at every cut-point: the F2, F3 and F5 deficits dominate, and the USA is closer to structured exposure than the four other jurisdictions, under both the $\geq$2 and the score-3 definitions (structured-only collapse in the Methods and Supplementary Appendix H). We privilege the score-3 threshold not because narrative disclosure is valueless (a score-1 PDF can support an individual manual lookup) but because procurement comparison, post-market surveillance, cross-device research, and multi-jurisdictional monitoring require structured, queryable fields rather than per-record document reading. A hospital may manually read one PDF for one device; a health system cannot maintain a life-cycle monitoring function by manually reconciling thousands of narrative records across authorisation, recall, and adverse-event systems. The audit therefore distinguishes narrative public availability from public-data infrastructure.

3.5 Gap-type distribution with confidence

Figure 2 reports the gap-type distribution across the 94 unmet cells (cells scoring below 3), stratified by confidence. Of the 94 unmet cells, 41 were structuring gaps, 28 schema gaps, 17 publication–access gaps, 7 linkage gaps, and 1 population gap. High- or Medium-confidence publication–access gaps accounted for 12 of the 94 unmet cells, limiting the claim that the missing public-data capacity can be restored mainly

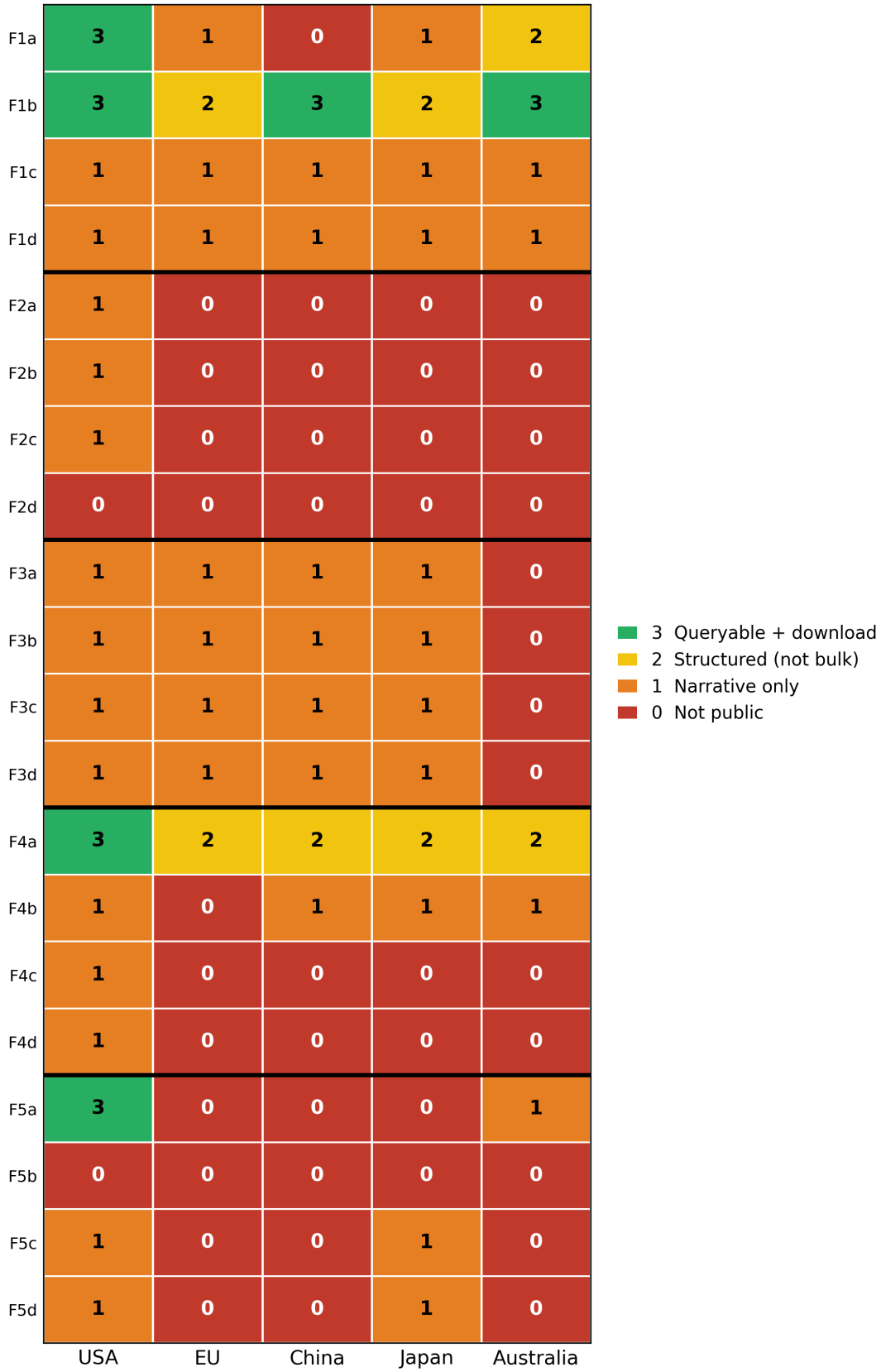


Figure 1: **Public availability of AI medical device requirements across 5 jurisdictions and 20 requirements (n = 100 cells, post-adjudication).** Colour gradient encodes the 0–3 availability score (red = 0 not public; orange = 1 narrative only; yellow = 2 structured but not bulk; green = 3 queryable and downloadable). The 20 requirements are grouped into five governance functions, separated visually by horizontal black lines: **F1 Identity** (F1a–d), **F2 Version & Change** (F2a–d), **F3 Evidence** (F3a–d), **F4 Linkage** (F4a–d), **F5 AI Specificity** (F5a–d). Cell-level confidence ratings are reported in Figure 2 (gap type stratified by confidence) and in the per-jurisdiction confidence breakdown in the Supplementary Information.

by publishing data already held by regulators. Strong reform claims tied to "regulators already hold the data" apply only to confirmed publication–access gaps and not to structuring or schema gaps.

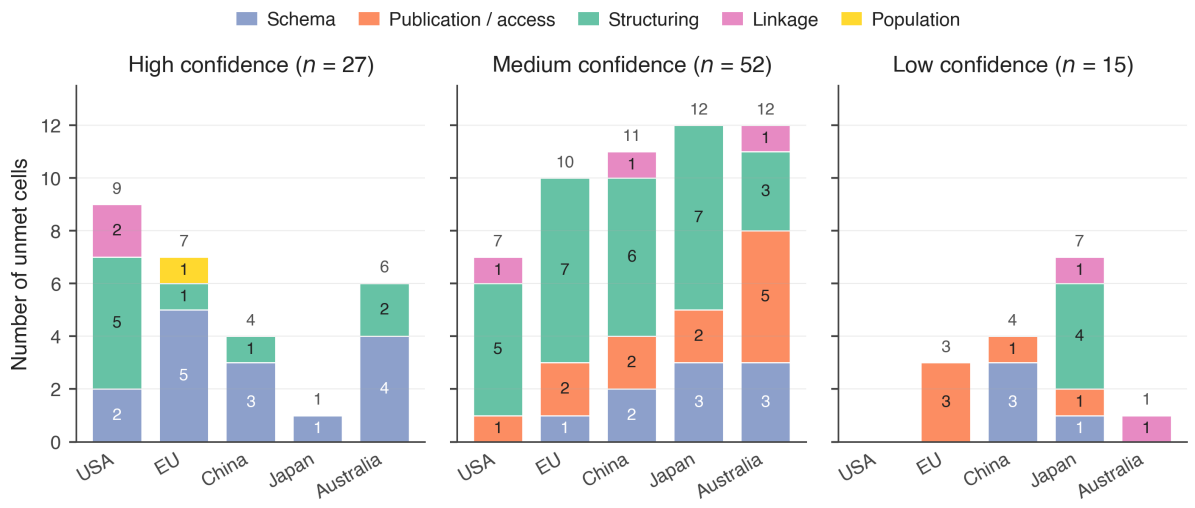


Figure 2: Five-way gap typology by jurisdiction, conditioned on per-cell confidence. Bars are stacked by gap type (schema / publication–access / structuring / linkage / population) and split into three confidence panels (High / Medium / Low, $n = 27$ / 52 / 15 unmet cells). Integer counts are printed on each segment and at the top of each bar, so the figure does not depend on colour alone. Generated from the post-adjudication dataset.

We report publication–access gaps stratified by confidence rather than as a single combined percentage, because only the High/Medium-confidence subset supports the claim that the underlying data is already held by the regulator and could be exposed without new collection.

3.6 Lifecycle rhetoric versus version-and-change traceability

Every jurisdiction articulates a lifecycle or change-management posture (the FDA PCCP framework, the EUDAMED "living picture", NMPA UDI updates, PMDA partial-change approvals, TGA software-change guidance), yet this rhetoric is matched by almost no public version-and-change traceability: across F2a–d, only the USA exposes any availability and only at the narrative/inferential level (PCCP guidance and 510(k) supplements), while the EU, China, Japan, and Australia expose none. A jurisdiction-by-jurisdiction summary mapping each lifecycle instrument to the audited F2a–d availability is in Supplementary Appendix J.

The probes are deliberately asymmetric because the public data environments are asymmetric: where bulk or API access exists we test recoverability programmatically, and where only search interfaces or documentation exist the architecture audit carries the comparison. The three stress tests reported here are therefore not parallel measurements of the same quantity across all five jurisdictions; they are jurisdiction-specific probes designed to make distinct audit claims observable in the public data environments where they can be tested programmatically. The FDA identifiability test asks whether known AI devices can be recovered from primary FDA pathway records without the curated AI list; the NMPA UDI test asks whether a bulk-downloadable product-identifier infrastructure supports structured AI findability; the FDA product-family test asks whether the richest available API environment exposes structured lineage for multi-submission AI product families. The comparative audit is carried by the 100-cell architecture matrix; these probes target selected failure modes rather than attempting a symmetric cross-jurisdictional comparison.

3.7 Targeted probe 1: FDA primary-record AI-status findability versus the curated AI list

This probe asks whether known AI devices can be recovered from the FDA's primary records (510(k), PMA, De Novo) without the dedicated AI/ML-enabled device list. The list is itself the FDA's AI identification mechanism; the probe measures whether AI status is also exposed as a structured field within the primary databases, not whether the FDA can identify AI devices at all.

The FDA AI-list CSV contained 1,430 entries (1,376 510(k), 17 PMA, 37 De Novo). The quantitative test is stratified by pathway and run against openFDA bulk endpoints, covering the 1,393 510(k)+PMA entries (97.4% of the list); the 37 De Novo entries are reported as a protocol constraint (openFDA exposes no bulk De Novo endpoint at cut-off; the identifier list is deposited in Supplementary Appendix E) rather than a public-record architecture finding. Full field lists and the tiered keyword set are in Methods.

For the 510(k) pathway (n = 1,376), each K-number was matched into the openFDA bulk 510(k) dataset and its primary searchable fields screened against the tiered AI keyword list. **Only 35 records (2.5%) carried a narrow AI/ML term in any primary field; 117 (8.5%) carried any AI signal; 566 (41.1%) matched only broad function terms (e.g. "automated", "algorithm") not specific to AI; and 693 (50.4%) carried no AI signal whatsoever.** False negatives are dominated by AI devices with generic product descriptions (e.g. "BriefCase", "Image Processing Software").

For the PMA pathway (n = 17, 100% matched), **6 (35.3%) carried a narrow AI term (5 via the "CAD" computer-aided-detection family) and 6 (35.3%) carried no AI signal** (per-record classifications in Supplementary Appendix E). CAD/CADe/CADx counted as narrow only where record context indicated computer-aided detection; the "iCAD" manufacturer substring did not. The higher PMA narrow rate reflects concentration in mammographic CAD devices, a pathway-composition effect rather than uniformly more transparent disclosure. **Pooled (n = 1,393), 41 (2.9%) carry a narrow term and 699 (50.2%) carry no AI signal**: the 510(k) headline is unchanged by the PMA extension.

3.8 Targeted probe 2: NMPA UDI findability (bulk-downloadable, no structured AI status)

Because no official public ground-truth list of AI-enabled medical devices is available for China, we report findability under plausible search strategies rather than recall against a known population. China has three distinct public-facing record environments: (i) the NMPA UDI database, a bulk-downloadable product-identifier dataset with a full release at the audit cut-off (`UDID_FULL_RELEASE_20260501.zip`; MD5-verified; ≈5.68M records across 1,136 XML parts; reconciliation notes in Methods and Supplementary Appendix A); (ii) the registration-certificate database, the more direct approval-record environment, exposed only through query interfaces; and (iii) technical review reports, narrative and not systematically downloadable. The probe is therefore not a complete audit of Chinese approval evidence; it is a findability test against the only bulk product-level source, asking whether that infrastructure also exposes AI status as structured fields. (The separate `nmpa.gov.cn/datasearch/` front-end is query-only, functionally distinct from the bulk release.)

The China probe shows a different failure mode from the FDA probe: not a lack of bulk data, but bulk data whose schema is not AI-aware. Applying the pre-specified bilingual keyword protocol (Section 5) to the free-text product-description field (`cpms`) yielded 119 records (0.002%) on narrow Chinese AI/ML terms, 76,435 (1.346%) on broad function terms dominated by the broad "assistive/auxiliary" term (*fuzhu*), and 1,744 (0.031%) on English/abbreviation terms; the deduplicated union is 77,304 records (1.361%). The broad tier was intentionally inclusive and carries a high expected false-positive rate (the *fuzhu* "assistive/auxiliary" term matches generic "assistive" usage); its stratified 300-record sample is

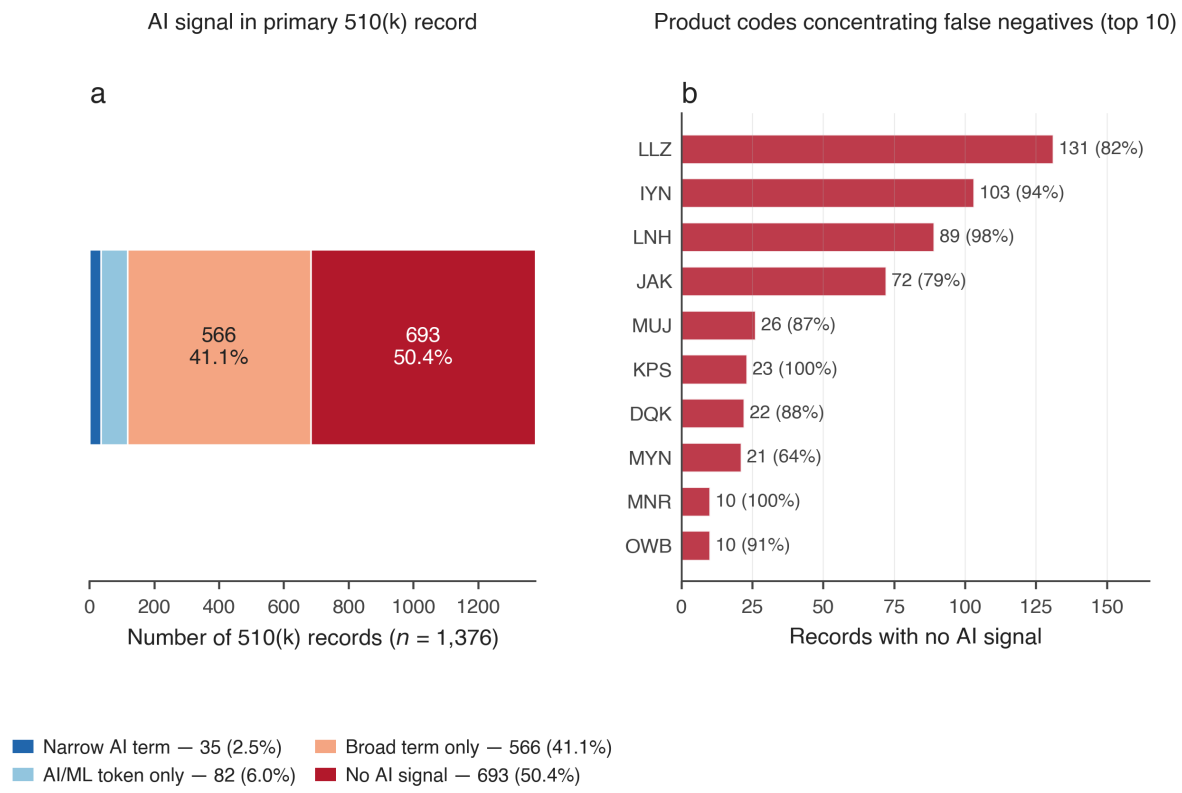


Figure 3: **FDA 510(k)-pathway AI identifiability sub-test** (n = 1,376 510(k) entries on the FDA AI/ML-Enabled Device List). **a**, Each device is classified by the strongest AI signal in any primary 510(k) field (`device_name`, `applicant`, `statement_or_summary`, `product_code`, `advisory_committee_description`, `openfda.device_name`; openFDA bulk dataset): only 35 of 1,376 records (2.5%) carry a narrow AI/ML term and 693 (50.4%) carry no AI signal at all. **b**, The ten product codes that concentrate the no-AI-signal records; the percentage is the share of that product code's AI-list devices carrying no AI signal in any primary field.

deposited as a transparency sample to document the retrieval burden and support future validation, and no prevalence or recall claim depends on it. These are keyword-retrieval counts, not prevalence estimates of AI devices in China; per-term and `flbm`-clustering breakdowns are in Supplementary Appendix A.

These findings indicate that NMPA UDI is publicly bulk-downloadable but does not expose AI status as a reliable structured field: the 70-tag UDI record schema contains no AI-status, AI-task, model-class, training-data-modality, or model-version field, and AI identification is possible only by free-text keyword inference against the descriptive `cpms` field. The structured `versionNumber`, `versionTime`, and `versionStauts` (sic) tags track the registry record itself, not the software / model version. This is consistent with the architecture audit (China F1b scores 3 on the bulk-downloadable structured UDI-DI, while F1a and F5a–d remain 0 with no AI-specificity field in the schema) and indicates that any AI-specificity reform in China would require adding structured AI metadata fields to the UDI record, not creating bulk-download access (which already exists).

3.9 Targeted probe 3: FDA product-family lineage (structured traceability in the richest API environment)

We selected six known multi-submission AI-device families as a purposive stress test of product-line traceability in the richest public data environment: Aidoc Medical (n=34 510(k) records), iCAD (n=15), HeartFlow (n=8), Caption Health (n=6), Cleerly (n=4), and Paige.ai (n=3). The selection is purposive, not a representative sample of all FDA AI devices; sampling intent and inclusion notes are in Supplementary Appendix F. For each record we tested whether predecessor / successor / product-family linkage is recoverable from a *structured* public field (vs. only from narrative text or string match on `device_name`).

Zero of 70 records expose a structured predecessor / predicate / related-510(k) field in the openFDA `/device/510k.json` schema. 7 of 70 carry an explicit version string in `device_name` (e.g. "Cleerly Labs v2.0"), enabling narrative inference for that minority. All 70 can be grouped by the `applicant` field, but applicant grouping conflates unrelated products and provides no temporal lineage direction. Predicate-device naming exists only inside the per-record 510(k) Summary PDF narrative, not as structured data. This does not directly measure the other four jurisdictions, but it shows that even the richest public API in the audit does not expose structured lineage for the sampled AI product families; the audit's F4d Schema-gap finding for FDA is supported empirically. Per-family timelines and the per-record openFDA dump are in Figure 4 and Supplementary Appendix F.

3.10 Clinical governance implications

The audit findings map onto four concrete clinical-governance tasks (Table 6). The mapping is constructed from the audit findings; it is not a set of patient-harm case studies and does not estimate clinical outcome effects. It identifies, instead, the public-data conditions required for safety assurance, procurement comparison, portability assessment, and safety-signal investigation.

3.11 A worked public-record traceability test

To make the audit concrete for a single clinical-governance case, we ran the five governance questions against the richest public record environment (the FDA) for a real AI-enabled device in our data, Aidoc's *BriefCase* triage family (34 510(k) records); this is a test of what a public user can recover, not an assessment of the device, and makes no safety claim. Even in this best case, three of the five questions cannot be answered from structured public records and a fourth only by manual PDF reading: AI status is recoverable only via the supplementary list (the primary `device_name` "BriefCase" carries no AI term), no structured current-version field exists, lineage is recoverable in 0/34 records as a structured

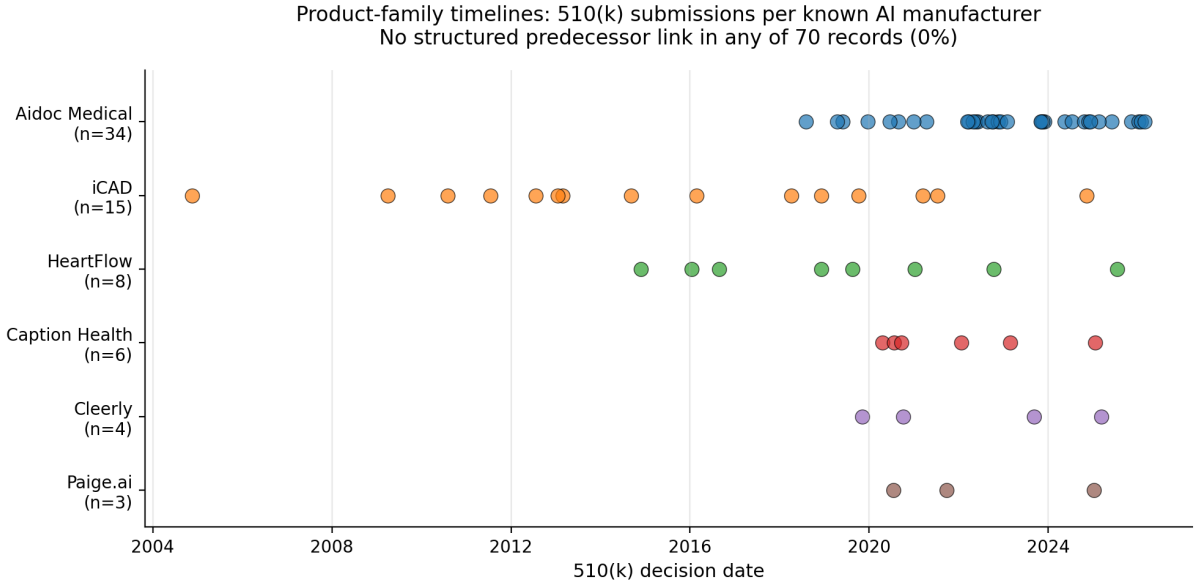


Figure 4: Product-family traceability for six known AI manufacturers (n=70 510(k) records, openFDA snapshot 2026-05-04). Each row is one manufacturer; markers are 510(k) decision dates. **Zero of 70 records expose a structured predecessor / predicate / related-510(k) field**; lineage is visible only as date-ordered submissions by the same applicant or as version strings in device names (10% of records).

Table 6: Clinical governance tasks for AI medical devices and the corresponding public-data audit findings. The table reports, for each task, what public data is needed, what the audit found, the practical consequence for clinical governance, and the requirement(s) addressing it.

Clinical governance task	**Public data needed**	**Audit finding**	**Practical consequence**	**Req(s)**
Procurement	Validation setting, validation population, performance metrics, stated limitations	F3b–d below structured threshold in all five jurisdictions; 12 of 15 F3b–d cells are structuring gaps and 3 are publication/access or schema gaps	Comparative review depends on manual PDF reading or, where summaries are not published, on regulator/vendor outreach	F3b–d
Software update	Current software version, public change history, regulatory basis for change	F2 cells score 0 in 17 of 20 cells (only narrative-level FDA evidence nonzero); FDA product-family traceability test shows no structured lineage across 70 records	Local clinical safety case cannot be updated from public records after a manufacturer change	F2a–c, F4d
Portability and equity	Validation population, data modality, training-data provenance	F3c, F5c, F5d are absent or narrative across all five jurisdictions	Distribution-shift and equity assessment cannot be performed at scale using public records	F3c, F5c–d
Safety-signal investigation	Stable device identifier linking authorisation, recall, and adverse-event records	F4b–c are linkage gaps in 6 of 10 cells; data exists separately but not joinable to the device record by a stable key	Safety-signal investigation requires manual cross-source reconciliation across separate public systems	F1b, F4b–c

field, validation population is narrative-only, and recall/adverse-event links require manual cross-system reconciliation. The full question-by-question table is in Supplementary Appendix I; the other four jurisdictions would return fewer structured answers still, their AI-status, version, and evidence fields scoring at or below the FDA's (Table 5).

3.12 Exploratory: jurisdictional similarity

Pairwise Jaccard similarity on binarised gap profiles is reported in Supplementary Appendix G as an exploratory analysis; we do not draw causal claims about the role of regulatory tradition in shaping public regulatory record environments from a five-jurisdiction sample with hand-coded requirements.

4 Discussion

Our findings describe a structural mismatch with direct clinical-governance consequences. Regulatory philosophy has moved toward lifecycle oversight of AI medical devices, but the public records through which hospitals, researchers, and patient-safety actors encounter those devices remain registration-oriented. They are *registration* infrastructures, organised around the one-time question of whether a device is legally marketed, rather than *governance* infrastructures able to support identification, change-tracking, evidence appraisal, post-market linkage, and AI characterisation after a device reaches the clinic. The audited environments can establish that an AI device entered the market, but they cannot, through public records, establish that it can be governed once there. The five-function specification names what the second capability would require; the audit measures how far the first falls short of it.

The contribution is therefore a source-grounded specification together with a transparent audit of how far the public regulatory record environments examined fall short of it: a derivation that is auditable through the four-anchor traceability matrix, an audit reproducible from the source inventory and codebook, and findings stable under the conservative core specification and the high-confidence cell subset.

The five functions are derived as a candidate minimal set for the governance tasks specified here, each independently anchored in regulatory, governance, clinical/operational and empirical/problem evidence (Section 2 and the traceability matrix). Dropping any one leaves a governance capability that the audited public regulatory record environments visibly fail to provide: F1 without F2 gives identity without currency; F3 without F4 gives evidence without outcomes; F1 through F4 without F5 describes a device but cannot express what makes it AI.

Because the five-way gap typology distinguishes structurally different reform targets, reform recommendations must be tiered. We adopt the following language discipline.

A confirmed publication–access gap (the regulator holds the data and could publish it without new collection) supports language such as "this could be addressed through publication of information already required or reviewed". A probable publication–access gap (public guidance suggests the regulator may hold the data but the public record does not allow confirmation) supports only the qualified form "this appears likely to require publication of information already held". A structuring gap supports "this requires structuring of information that is already disclosed in narrative form", not a transparency mandate. A linkage gap supports "this requires identifier infrastructure connecting existing public sources", not a transparency mandate either. A schema gap supports "this requires new structured fields or submission requirements". A Low-confidence gap of any type supports only an indicative claim, never a strong reform conclusion.

We do not make the blanket claim that the largest share of unmet cells is publication–access gaps and that transparency mandates are therefore the single highest-impact reform. The five-way classification,

when stratified by confidence, distributes unmet cells across all five gap types, and the relative share of confirmed publication–access gaps is reported in Section 3 rather than asserted as a headline finding.

F5 is the worst-served function and the most distinctive finding. The deficit is one of integration, not total absence: the FDA AI-Enabled Device List and the TGA AI list expose AI-specific information through supplementary agency-level catalogues, but it is not integrated into primary registry architecture as structured, queryable, lifecycle-linked metadata. F1–F4 deficits are deficits of degree (the environments identify, track, evidence, and link devices, but incompletely or in unstructured form), whereas F5b–d are deficits of *kind*: the categorical vocabulary needed to describe AI devices *as AI devices* is not exposed as structured fields. This is not AI exceptionalism: F1–F4 apply to software-intensive devices generally and their versioning problems are shared with non-AI software; F5 isolates only the additional metadata needed where behaviour depends on model class, training/validation data, and post-deployment adaptation.

This matters because AI-specific governance depends on AI-specific information. Post-market surveillance for an adaptive or continuously learning device requires knowing that the device adapts; equity assessment requires knowing the modality and provenance of the training and validation data; comparative effectiveness research benefits from knowing the model/method class. None of these are possible at scale when the public record exposes a device name and a product code but no AI-specific metadata.

The EU AI Act [European Parliament, 2024] introduces a European AI database that, for medical devices, is separate from EUDAMED. For medical AI, separate AI and medical-device databases risk splitting the device identity from the AI metadata needed to monitor it: an AI medical device would appear in two regulatory data environments without a guaranteed cross-reference. The relevant design principle is therefore identifier linkage and field-level integration of AI-specific metadata into the device's public regulatory record, not the creation of another standalone disclosure portal.

We propose a concrete public-data extension of 20 structured fields (one per audited requirement) that, if added to existing device registration records, would close the specification gaps the audit identified. The proposed extension is intentionally modest: it asks for categorical summary metadata sufficient to identify, compare, trace, and link AI medical devices, and it does not require publication of source code, model weights, training datasets, patient-level data, or detailed proprietary architecture. Table 7 gives the full artifact: for each field, the requirement it satisfies, its data type, whether it can be populated from a controlled vocabulary, its public-governance value, the gap type it addresses, its disclosure risk, and a concrete implementation route. The table is designed to be liftable directly into a regulator or standards-body consultation document.

Each row of Table 7 is anchored in a requirement that is itself anchored in the four-anchor traceability matrix; we propose no field without a corresponding requirement. The 16 unmarked rows form the conservative claim-bearing set; the four † rows (F4d product-family linkage and F5b–d AI-technical metadata) carry Medium inclusion confidence and are the extended edge of the proposal. Truly out-of-scope items that have no audited requirement (training-dataset size, equity-stratified subgroup performance, cybersecurity metadata) remain future extensions and are deliberately excluded even here.

The extension calls for structured summary metadata, not raw-data disclosure. It does not require publication of training datasets, model weights, source code, or detailed proprietary architectures, and so does not turn on releasing trade secrets or patient-level data. Most of the proposed fields can be populated from controlled categories rather than free text: model/method class as a small enumeration (e.g. deep learning / classical machine learning / rules-based / hybrid / foundation-model-enabled); training and validation data modality as a controlled list (e.g. CT / MRI / X-ray / EHR / pathology / waveform); validation geography as single-country / multi-country / unspecified; and update status as locked / periodically updated / adaptive / continuous-learning. Disclosure risk is therefore field-specific: several fields can be published directly, others as categories, and a minority (for example, detailed

Table 7: Candidate minimum public data extension: 20 structured fields mapped one-to-one onto the audited requirements. "Ctrl. vocab" = field can be populated from a controlled enumeration rather than free text. "Disclosure risk" is field-specific and refers to summary metadata, not raw data (no training datasets, weights, source code, or patient-level records). Rows marked † (F4d, F5b–d) carry Medium inclusion confidence and lie at the extended edge of the specification; the other 16 form the conservative claim-bearing set.

Field	Req.	Type	Ctrl. vocab	Public value	Gap addressed	Discl. risk	Implementation route
`ai_status`	F1a	flag	Yes	Identify AI/software devices	Schema	None	New enum field
`device_id`	F1b	identifier	UDI-DI	Cross-system linkage	Linkage	None	Reuse UDI-DI
`intended_use`	F1c	text	Partial	Clinical scope	Structuring	None	Structure existing text
`function_code`	F1d	code	Yes	Comparison / HTA	Structuring	None	New taxonomy field
`sw_version`	F2a	string	semver	Deployed version known	Schema	None	New vendor-supplied field
`change_log`	F2b	link/log	n/a	Trace modifications	Pub./linkage	Low	Publish change log
`change_basis`	F2c	enum	Yes	Authority for change	Schema	None	Enum: 510k/PCCP/suppl.
`update_status`	F2d	enum	Yes	Adaptivity exposed	Schema	None	locked/periodic/adaptive
`evidence_url`	F3a	link	n/a	Evidence access	Pub. access	Low	Publish decision summary
`val_setting`	F3b	enum	Yes	Portability	Schema	Low	New field
`val_pop`	F3c	summary	Partial	Equity / portability	Schema	Low–med	Summary, not patient data
`perf_summary`	F3d	structured	metrics	Evidence appraisal	Structuring	Low	Structure AUC/sens/spec
`auth_status`	F4a	enum	Yes	Current standing	(present)	None	Reuse certificate field
`recall_key`	F4b	key	ID	Connect recalls	Linkage	None	Key infrastructure
`ae_key`	F4c	key	ID	Connect AE reports	Linkage	None	Key infrastructure
`family_key`†	F4d	key	ID	Device lineage	Linkage	None	Predecessor/successor key
`ai_task`	F5a	enum	IMDRF	What the AI does	Schema	None	New enum (IMDRF terms)
`method_class`†	F5b	enum	Yes	Surveillance design	Schema	Low	DL/classical/rules/hybrid
`train_modality`†	F5c	enum	Yes	Portability / shift	Schema	Low	CT/MRI/EHR/path/waveform
`train_prov`†	F5d	summary	Partial	Shift assessment	Schema	Med	Summary; delayed/partial ok

provenance) may warrant summary-level or delayed publication. Framing the extension as categorical summary metadata is what makes it implementable without exposing confidential technical detail.

Table 8 presents six priority recommendations, ordered by feasibility and impact.

The recommendations are ordered by gap type rather than by jurisdiction, on the rationale that the publication–access vs structuring vs linkage vs schema distinction determines the kind of policy instrument required (transparency step / structuring step / identifier infrastructure / new fields and submission requirements). Recommendations are not made for Low-confidence cells; those are reported in Supplementary Appendix D as indicative.

Where prior work identifies recurring database problems, this paper turns those problems into an auditable public-data specification and a reform typology; the distinction matters because a missing field, an unpublished field, an unstructured PDF disclosure, and an unlinked identifier require different institutional remedies. Relative to the four challenges identified by Ong et al. [Ong et al., 2026], the present audit differs in three respects.

First, the source-grounded specification moves from problem description to a transparently derived requirement set. Ong et al. characterise problems; the specification here proposes minimum requirements with explicit anchors (regulatory, governance, clinical/operational, empirical/problem) and explicit inclusion confidence ratings.

Second, the five-way gap typology refines the older three-way scheme. Ong et al.'s challenges (inconsistent identification, limited version tracking, insufficient transparency, weak linkage) map onto distinct gap categories in our typology; in particular, "insufficient transparency" corresponds to a mix of publication–access, structuring and linkage gaps that the earlier scheme would have collapsed. The distinction matters because it determines whether the right policy instrument is transparency regulation, structuring of already-disclosed narrative content, identifier infrastructure, or new fields.

Third, F5 (AI Specificity) is introduced as a governance function with explicit anchors. Within the five jurisdictions audited here, F5b–d are absent as structured fields, indicating an absence at the level of structured public registry fields rather than a data-quality shortfall, and explaining why prior frameworks built around generic device-registry problems did not surface them. Reporting standards for AI prediction models such as TRIPOD-AI [Collins et al., 2024] and the early-stage clinical evaluation guideline DECIDE-AI [Vasey et al., 2022] prescribe what individual studies should disclose, and align with several of our F3 and F5 requirements; the audit here is complementary, characterising what *public regulatory record environments* expose at scale, rather than what individual studies should report.

Several limitations qualify these findings. First, the specification is source-grounded, not externally validated. We did not run an expert panel, Delphi process, or formal validation exercise. The four-anchor traceability matrix and the two sensitivity restrictions (core-only, High/Medium-confidence) substitute transparent derivation and stability testing for external validation, but they do not establish expert consensus. Both annotators were also study authors and so share the project's interpretive priors; to mitigate this, coding was independent, source-locked, and rule-governed, with every adjudicated change recorded in a public log. External validation of the audit instrument (an independent third annotator on a stratified subsample and a structured stakeholder review of the 20 requirements) remains the most valuable next step and is future work.

Second, the audit examines public regulatory record environments, not internal regulatory capacity. A publication–access gap is consistent with the regulator collecting and reviewing the relevant information internally; we do not claim that it does not. Reform claims tied to "regulators already hold the data" apply only to High- or Medium-confidence publication–access gaps, and we report those separately rather than as a headline figure.

Third, the audit is a cross-sectional snapshot at the cut-off, and evidence depth is uneven across jurisdictions. EUDAMED, the FDA PCCP framework, and Australia's (under one year old) AI list are

Table 8: Priority recommendations tiered by gap type and inclusion confidence. Language tier (strong / moderate / cautious / indicative) is assigned by the rule in Section 5. Recommendations addressing High- or Medium-confidence cells use stronger language; those addressing Low-confidence cells are framed as indicative.

Requirement(s)	**Gap type**	**Recommendation**	**Language tier**	**Actor**
F3a, F3d (where confirmed)	Publication–access	Publish structured plain-language decision and performance summaries for AI device authorisations, on the FDA pattern of 510(k) Summaries / Statements	Strong (where High-conf publication–access) / Moderate (Medium)	NMPA, PMDA, TGA
F3b, F3c, F3d (where narrative-public)	Structuring	Convert narrative validation and performance information into structured fields with controlled vocabularies	Strong	FDA/CDRH; NMPA/CMDE; PMDA; TGA
F4b, F4c (where unjoined)	Linkage	Use UDI-DI as the cross-source key between the primary regulatory record, recall, and adverse-event databases	Strong (where joinable in principle) / Moderate (where identifier coverage incomplete)	FDA/CDRH; DG SANTE/EUDAMED; NMPA; PMDA; TGA
F1b (where field exists, sparse)	Population	Submission validation rules requiring stable identifier population	Moderate	FDA/CDRH; DG SANTE/EUDAMED; NMPA; PMDA; TGA
F5a (where currently inferred or list-based only)	Schema or Publication–access	Add a structured AI-task-category field within the primary public regulatory record, not a parallel list	Strong (where confirmed schema gap) / Moderate (where probable publication–access)	IMDRF AIMD WG; FDA/CDRH; DG SANTE/EUDAMED; NMPA; PMDA; TGA
F5b, F5c, F5d	Schema	Add structured fields for model/method class, training/validation data modality, and training/validation data provenance summary; align through IMDRF AI key terms	Strong on schema-gap claim; cautious on patient-impact framing	IMDRF AIMD WG (terminology); FDA/CDRH, DG SANTE/EUDAMED, NMPA, PMDA, TGA (implementation)
F2a, F2b, F2c, F2d	Schema (with publication–access component for some jurisdictions)	Add structured fields for current software/model version, public change/update history, regulatory basis for change, and update status (locked / periodically updated / adaptive / continuous-learning)	Strong on schema-gap claim; moderate on transparency-mandate component	FDA/CDRH; DG SANTE/EUDAMED; NMPA; PMDA; TGA

all evolving; the EU column should be read as a pre-mandatory-use baseline for the first four EUDAMED modules, whose mandatory use began on 28 May 2026 (27 days after our cut-off) under Commission Decision (EU) 2025/2371 (Vigilance is targeted for Q2 2027). Source depth ranges from bulk-downloadable structured data (FDA via openFDA; the NMPA UDI 2026-05-01 release, MD5-verified) through search-interface-only data to documentation-only sources. This asymmetry is propagated into the per-cell confidence rating, and the High/Medium-confidence sensitivity analysis re-runs the main findings on the subset where evidence depth is highest. The findings are a baseline against which progress can be measured, not a permanent characterisation.

Fourth, gap-type assignment involves judgment, particularly at the schema-vs-publication-access and structuring-vs-publication-access boundaries; the codebook documents single-cause rules and the adjudication log records every changed cell. Confidence rating was the least reliable of the three coded dimensions (raw 0.740, κ 0.555, cautious), which is exactly why per-cell confidence is used only to *constrain* claim strength (strong recommendations restricted to High/Medium-confidence cells) rather than to generate any primary finding.

Finally, the audit is bounded in scope in three ways. It covers five jurisdictions, with South Korea, Canada, and Brazil notable omissions that the instrument is designed to accommodate but that we did not code. Two probes are deliberately non-symmetric. The China findability probe (Section 3.8) is a keyword-retrieval test against the UDI `cpms` field, not a recall or prevalence estimate; the broad tier was intentionally inclusive and has a high expected false-positive rate, and the deposited stratified sample is provided to document the retrieval burden and support future validation, but no prevalence or recall claim depends on it. The FDA product-family probe was run against openFDA only, establishing the FDA result directly (0% structured lineage in 70 records) but only motivating the cross-jurisdictional concern. And equity-stratified performance, real-world performance, cybersecurity metadata, deployment monitoring, human-factors evidence, and full training-dataset documentation sit in the unscored extended specification, beyond the minimum public-data extension proposed here.

Reform recommendations are tiered by gap type and confidence rather than presented as a single headline reform. A confirmed publication–access gap supports a transparency step; a structuring gap supports structuring of already-disclosed narrative content; a linkage gap supports identifier infrastructure; a schema gap supports new fields and submission requirements. The candidate minimum public data extension proposed here maps each new field directly onto an audited requirement, so that no field is included without a justification traceable through the matrix.

Lifecycle oversight of AI medical devices cannot depend only on internal regulatory review. If hospitals, researchers, and patient-safety actors are to evaluate, monitor, and contest these systems after authorisation, public records must expose stable device identity, software change, validation evidence, post-market linkage, and AI-specific metadata in structured form. Until they do, public registries can establish that an AI device entered the market, but not support its governance after entry.

The dated finding is that, as of the audit cut-off, five leading public regulatory record environments were organised around market-authorisation visibility rather than lifecycle governability. The durable contribution is the reusable apparatus: a public-governability framework, a 20-requirement audit instrument with a defensible 12-item core, a five-way gap typology that maps deficits to distinct reform instruments, and a candidate minimum public-data extension. Together these provide a baseline against which future registry development, cross-jurisdictional regulatory benchmarking, and standards work can be measured as public records evolve.

5 Methods

5.1 Study design

This is a two-annotator, source-grounded audit of the public regulatory record environments of five jurisdictions against a 20-requirement specification of governance-supporting public data infrastructure for AI medical devices. The specification is derived through a four-anchor source-concordance procedure (Section 2) and is not subjected to expert-panel validation: the purpose of the study was not to produce an expert-consensus standard, but to test whether a transparently derived, source-grounded public-data specification could be applied reproducibly to existing public regulatory record environments. The audit reports findings under a 12-item core specification (used for conservative-sensitivity analysis) and the full 20-item set, supplemented with two sensitivity restrictions and three empirical stress tests. We do not claim that the specification represents expert consensus or that the audit "validates" the specification; we claim that the derivation is transparent, that the audit is reproducible, and that the patterns it surfaces are stable under the conservative core specification.

5.2 Unit of analysis

The unit of analysis is the **public regulatory record environment** of each jurisdiction, defined as the set of official, publicly accessible regulatory data sources through which a user can identify, evaluate, trace or monitor a medical device without privileged regulatory access. For each jurisdiction this comprises the primary device registry; any AI-specific device list; the UDI database; public review or decision summaries; certificate or authorisation status records; recall or safety notice databases; adverse-event or incident reporting databases; and official data dictionaries, field specifications, public APIs and regulatory guidance. This unit is wider than "the registry" because for several jurisdictions (notably the FDA, EUDAMED, PMDA and the TGA) the relevant public information is split across several official systems, none of which alone constitutes the registry.

5.3 Jurisdiction selection and evidence depth

We selected five jurisdictions representing the major regulatory traditions for medical devices: **China (NMPA)**, **USA (FDA)**, **EU (EUDAMED)**, **Japan (PMDA)**, and **Australia (TGA)**. Selection covers the three largest device markets (US, EU, China), two significant Asia-Pacific regulators (Japan, Australia), and spans pre-market notification (US 510(k)), conformity assessment (EU, Australia), and centralised approval (China, Japan). For each jurisdiction we record an *evidence-depth* classification: bulk-downloadable structured data with directly observable fields (FDA 510(k), PMA, De Novo, MAUDE, FDA Recalls, all distributed via openFDA, the FDA AI-Enabled Device List CSV, and the NMPA UDI 2026-05-01 full release); structured data accessible only via search interface (PMDA Shonin; ARTG; EUDAMED for the four first-cohort modules at the cut-off); and documentation-only sources (EUDAMED Vigilance and Market Surveillance modules at the audit cut-off, the EU mandatory-use timeline being sourced to Commission Decision (EU) 2025/2371; TGA SaMD guidance and AI-enabled-devices guidance accessed as Akamai-protected pages where HTML snapshot was unobtainable). Evidence-depth is reported per source and propagated into per-cell confidence ratings (Section 5).

5.4 Source-derived specification

The 20-requirement specification is derived through a four-anchor source-concordance procedure: each requirement is checked against (i) a regulatory anchor (concept named in official guidance, regulatory database fields, submission requirements, UDI structure, PCCP logic, SaMD framework, AI-device list,

or equivalent official source); (ii) a governance anchor (concept needed for identification, comparison, tracing, linkage or AI characterisation); (iii) a clinical/operational anchor (concept needed for procurement, deployment, local validation, clinical safety case review, audit, or post-market monitoring); and (iv) an empirical/problem anchor (prior literature or our own probing showing that absence creates a public-data problem). A requirement is retained in the full specification if it satisfies at least two anchors, with inclusion confidence rated High (regulatory and governance both present), Medium (regulatory or governance present plus at least one other), or Lower (clinical/operational and empirical/problem only). The 12-item core specification is the subset for which both regulatory and governance anchors are present. The full traceability matrix is in the supplementary material.

5.5 Data collection and source inventory

For each jurisdiction and each source in the public regulatory record environment, we recorded the source name and type, the access URL, the access mode (API / bulk download / search interface / PDF / HTML), the retrieval date, the archive status (locally saved / screenshot captured / accessed live), the available fields, searchability, downloadability, language, translation method (manual / machine / bilingual verification), and any limitations. The complete source inventory is provided as a supplementary appendix. For non-English sources (Chinese NMPA, Japanese PMDA), translations were produced by a combination of machine translation and bilingual verification by Annotator A; the translation method is recorded per cell. Sources for which the link was unreachable, behind a login wall, or geographically restricted at the time of access are flagged in the inventory and excluded from coding. A pre-submission source-currency re-check (2026-06-16) re-resolved all 24 sources and is deposited as `data/sources/pre_submission_source_check_20260616.csv`; the only material change was the FDA AI-Enabled Device List page relocating to a new URL (the list itself unchanged), with several PMDA search endpoints also relocated and the TGA pages remaining CDN-restricted. No re-checked change altered a scored cell: the audit is a cross-sectional snapshot locked at the 2026-05-01 archive, and post-cut-off changes are treated as context rather than used to silently revise the locked matrix.

5.6 Coding framework

For each (jurisdiction, requirement) cell, two annotators independently recorded four values: *availability score* on a 0–3 scale (0 = not publicly available; 1 = public only in narrative, PDF, free-text, manual or inferential form; 2 = public as a structured field but not queryable / downloadable / consistently accessible; 3 = public as a structured, queryable or downloadable field); *gap type* when score < 3, selected from five categories (schema; publication–access; structuring; linkage; population); *confidence rating* (High = direct evidence from API, bulk download, field list, registry interface, official data dictionary, or observed public record; Medium = supported by official documentation but not directly observed in structured data; Low = inferred from ambiguous documentation, translation uncertainty, inconsistent interface behaviour, or incomplete public sources); and an *evidence note* citing the specific source consulted, the exact field name or section, the retrieval date, and the translation method where applicable. The full codebook (including boundary rules, single-cause assignment for gap type, and the population-fill cap that limits a sparsely populated field to score 2) is provided as supplementary material. A field with the right name but the wrong tracked entity (e.g., a registry-record version field used in place of a software version field) is scored against the requirement, not the field, and the mismatch is documented in the evidence note.

Both annotators read the codebook and the source inventory before coding. The codebook went through four locked versions (v0 $\rightarrow$ v0.1 $\rightarrow$ v0.2 $\rightarrow$ v1), each rule change documented in the changelog and tied to a specific reviewer comment or pilot-surfaced ambiguity. A pilot round of 20 deliberately

difficult cells, covering all five functions, five gap-type boundaries, and three score-boundary tests, was coded independently under a 2-day split with no contact during coding. Pre-adjudication pilot agreement was raw 0.850, Cohen's κ 0.831 on the availability score (Strong), with zero two-step disagreements. The three pilot disagreements were resolved by email-based adjudication into three rule clarifications added to codebook v1: a non-operational-register clarification (§2.5), a F1a inferential-identification rule excluding broad-keyword scoring (§2.6), and a Wayback-substitution score cap at 2 (§4.2).

The full 100 cells were then coded independently by both annotators against codebook v1, with no discussion until both had submitted, and against a source base locked at the 2026-05-01 archive (24 sources). Pre-adjudication agreement on the full coding was raw 0.780, linear-weighted Cohen's κ 0.724 on the availability score (Acceptable). Two two-step availability disagreements were identified (EU F2a, where Yu coded 2 against Aaron's 0; and EU F4d, where Yu coded 0 against Aaron's 2) and both were resolved through documented adjudication, the EU F2a disagreement under the Section 6 mismatch rule and the EU F4d disagreement under the spec Section 5 governance-probe note. No further codebook change was required. The 22 pre-adjudication disagreements clustered around four boundary problems: (i) the 1-vs-2 boundary on whether free-text intended-use content qualifies as "structured" under codebook Section 2; (ii) the 0-vs-1 boundary on whether narrative content was actually located vs. assumed; (iii) the F5c/F5d question of whether modality / provenance language in 510(k) Summary or PMDA review-report narratives suffices for score 1; and (iv) one Section 6 mismatch case (EU F2a UDI version field). Adjudication proceeded by email exchange under the rule that every change be traced to a specific codebook section; the adjudication log records, per changed cell, both annotators' codes, the final code, the codebook section invoked, and the decisive evidence source. Of 22 disagreements, 18 resolved with one annotator accepting the other's reading after re-application of the codebook, and 4 with the higher-coding annotator revising downward to the more conservative interpretation. The adjudicated dataset is the analysis-ready set used in Results.

For the ordinal availability score we report raw agreement and linear-weighted Cohen's κ (pre-specified). For gap type we report raw agreement and Cohen's κ, supplemented with Gwet's AC1 where category prevalence in either annotator's coding exceeded 60%, on the rationale that AC1 is less sensitive than κ to the high-agreement / high-prevalence paradox that arises when one gap type dominates; AC1 was not triggered in our coding. For confidence rating we report raw agreement and Cohen's κ. We use interpretation thresholds of ≥ 0.80 strong, 0.67–0.79 acceptable, 0.50–0.66 cautious, < 0.50 codebook failure for that dimension; per-dimension thresholds are reported even when one dimension fails.

Per-function reliability is reported in the Supplementary Information. F3 (Evidence Legibility) showed high raw agreement but unstable κ because most cells clustered around adjacent 0/1 categories and the disagreements concerned the boundary between "not located" and "narrative evidence located". We therefore interpret F3 results cautiously and support them with cell-level evidence notes; the principal F3 claim, which is limited to the absence of structured/queryable evidence fields, was stable after adjudication. The pre-adjudication coding sheets, the adjudication log, and codebook v0 through v1 are deposited (see Data Availability).

After adjudication, the main findings were re-run under two pre-specified restrictions to test their stability. *Core-only*: scores recomputed against the 12-item core specification (5 jurisdictions $\times$ 12 requirements = 60 cells), to test whether the main findings hold under the most conservative version of the specification. *High/Medium-confidence*: the main confidence restriction excluded Low-confidence cells and retained High- and Medium-confidence cells (n = 86), to test whether the findings depend on uncertain coding. A stricter *High-only* subset (n = 32) is reported descriptively in the Supplementary Information; the reduced cell count makes function-level comparisons unstable, so it is not used as a primary sensitivity restriction. A third pre-specified restriction (*structured-only*, scores 0–1 vs. 2–3 collapse) was run as a descriptive sensitivity check and is reported in Supplementary Appendix H. Under

the structured-only collapse, 13 of 100 cells (13%) score 2 or 3; the structured share is 0% in F2 and F3, 5% in F5, 25% in F4 and 35% in F1; by jurisdiction the structured share is 20% in the USA and 10–15% in the four other jurisdictions. The restriction does not alter the direction of the main finding (F2, F3 and F5 deficits dominate; the USA is closer to structured exposure than the four other jurisdictions). Because it largely duplicates the distinction already encoded in the 0–3 scale, the structured-only collapse is reported as a sensitivity check rather than as a primary inferential result. These sensitivity analyses are the primary substitute for external validation in the design adopted here.

5.7 Jurisdiction-specific stress tests

Three empirical stress tests give the audit observable form.

FDA AI identifiability test. Using the FDA AI/ML-Enabled Medical Device List as a reference set, we tested whether known AI devices can be recovered from the FDA's primary regulatory records (510(k), PMA, De Novo) without relying on the dedicated AI list. All entries on the live FDA AI-Enabled Device List were extracted (1,430 entries: 1,376 510(k), 17 PMA, 37 De Novo). The quantitative identifiability test is reported stratified by pathway because field names and disclosure structures differ. The 510(k) test uses openFDA `/device/510k.json` (n = 1,376); the PMA test uses openFDA `/device/pma.json` (n = 17). The 37 De Novo entries are reported as a methodology note because openFDA does not expose a bulk `/device/de_novo.json` endpoint at the audit cut-off; the De Novo identifier list is deposited (rows with submission numbers starting `DEN` in `data/fda/fda_ai_list_published_20260501.csv`). FDA De Novo records are publicly searchable and downloadable through a separate FDA De Novo database that was not incorporated into the present scripted openFDA protocol; we therefore report De Novo as a pathway-specific protocol limitation rather than a public-record architecture finding. Together the 510(k) and PMA tests cover 1,393 / 1,430 = 97.4% of the FDA AI list.

For the 510(k) sub-test, each K-number was matched into the openFDA bulk 510(k) dataset (174,817 records, official partition 0001-of-0001 of the openFDA download); match rate was 100%. For each matched record, primary searchable fields (`device_name`, `applicant`, `statement_or_summary`, `product_code`, `advisory_committee_description`, `openfda.device_name`) were screened against a tiered AI keyword list. For the PMA sub-test, the analogous primary fields are `trade_name`, `applicant`, `ao_statement` (the PMA analogue of `statement_or_summary`), `product_code`, `advisory_committee_description`, `openfda.device_name` and `generic_name`; match rate was also 100%. In both sub-tests each record was classified into one of four tiers: narrow-AI-term match (reliably identifiable as AI; *artificial intelligence*, *machine learning*, *deep learning*, *neural network*, or the CAD family), AI/ML token-only match (no narrow term but "AI" or "ML" present at word boundary), broad-term-only match (*automated*, *algorithm*, *smart*, *intelligent*, or *decision support*; insufficient for AI identification because of false-positive risk), or no AI signal. The PMA pipeline is deposited at `code/results_pipeline/regen_pma_identifiability.py` with the raw output at `data/fda/pma_identifiability_full_raw.csv`.

The framing is that the FDA's primary regulatory databases do not reliably expose AI status as a structured field, not that the FDA lacks an AI identification mechanism: the dedicated list is itself such a mechanism, and the test asks whether AI status is recoverable in its absence.

China findability analysis (pre-specified protocol applied to the NMPA UDI 2026-05-01 full release). Because no official public ground-truth list of AI-enabled medical devices is available for China, the NMPA analysis was pre-specified to estimate public findability under plausible search strategies rather than recall against a known AI-device population. The pre-specified protocol applied a locked bilingual keyword set, comprising 13 narrow Chinese AI/ML terms (English glosses: artificial intelligence, machine learning, deep learning, neural network, convolutional neural network, recurrent neural network, reinforcement learning, federated learning, self-supervised learning, transfer learning, generative, large

model, foundation model), 11 broader Chinese clinical/function terms (intelligent, computer-aided diagnosis, imaging-assisted, image processing, assistive/auxiliary, algorithm, prediction, automatic recognition, automatic detection, automatic segmentation, computer-aided), and 8 English/abbreviation terms (AI, ML, deep learning, neural network, algorithm, CAD, CADe, CADx), to the free-text product-description field (`cpms`) of the NMPA UDI `UDID_FULL_RELEASE_20260501.zip` bulk dataset (284.56 MB; MD5 `493615E48D67B3DC9D48CBC63301371F` matching the NMPA-published value; 1,136 XML parts). The exact Chinese-character forms of all 24 Chinese keyword terms are provided in the deposited locked bilingual keyword set (see Data Availability). The file header declared 5,676,167 records; the parse observed 5,678,685 `<device>` elements, of which 5,678,240 carried a unique `deviceRecordKey` (445 exact-key duplicates), leaving a residual of 2,073 elements above the header count attributable to a snapshot/timing difference between the header-declared database count and the exported XML. We use the header-declared count for denominator statements and report the parsed-element count as a data-quality note. Records matching at least one keyword in each tier were extracted; the union over the three tiers was deduplicated by `deviceRecordKey`. A product-classification-code (`flbm`) clustering check was performed on the deduplicated set. A stratified random sample of 300 broad-tier records (stratum = matched keyword) was pre-specified and is deposited as a transparency/verification sample for future human labelling; no prevalence or recall claim depends on it. Narrow and English hits are below the 200-record threshold and are deposited in full for future labelling. The processing script is deterministic; the keyword set is locked in the deposited methodology folder.

FDA product-family traceability test. We tested whether a public user can trace the evolution of an AI medical device product family through public regulatory records, restricted to the FDA on the rationale that the FDA is the richest public data environment and that traceability difficulty in the richest environment is informative. Six known AI-device families with multiple 510(k) submissions were sampled (Aidoc Medical, n=34; iCAD, n=15; HeartFlow, n=8; Caption Health, n=6; Cleerly, n=4; Paige.ai, n=3; total n=70 records), seeded from manufacturers known from the AI/ML list to have multi-submission histories. For each record we tested whether predecessor / successor / product-family linkage is recoverable from a *structured* public field, vs. only from narrative text or a string match on `device_name`. The pre-specified recoverability tiers were: structured predecessor / predicate / related-510(k) field exposed in openFDA; version string in `device_name`; same `applicant` field; or recoverable only via per-record 510(k) Summary PDF narrative. The cross-jurisdiction inference is conservative: at the audit cut-off the FDA was the only jurisdiction with a public programmatic API for 510(k)-equivalent records (openFDA), and the test was therefore run on the richest API in the audit. NMPA UDI is bulk-downloadable but does not expose product-family or predicate relationships as structured fields (verified post-cut-off against the 2026-05-01 full release schema: none of the 70 device-level tags is a predecessor / predicate / product-family field); EU EUDAMED, PMDA, and TGA do not expose a comparable bulk endpoint at the cut-off.

5.8 Clinical-governance vignettes

Four short analytic vignettes (procurement; software update; portability/equity; safety-signal investigation) were constructed after the audit findings closed, each citing a specific audit finding. Each vignette follows a fixed structure: clinical governance task; public data needed; audit finding; institutional failure created by the missing / unstructured / unlinked data; requirement(s) that would address it. We state explicitly that these are analytic clinical-governance scenarios constructed from audit findings; they are not patient-harm case studies and do not estimate clinical outcome effects.

5.9 Claim discipline

Because the unit of analysis is the public record environment, the claims are deliberately bounded; this restraint is a matter of scientific precision about what the evidence supports, not a hedge. Throughout the analysis we use three tiers of language matched to evidence basis. *Strong* language ("the public record does not expose…") is used only where directly supported by public evidence with High or Medium confidence. *Moderate* language ("this pattern is consistent with a structuring gap", "we classify this as a probable publication–access gap") is used where inference is involved. *Cautious* language ("the study cannot determine whether non-public regulatory systems contain more complete information") is used where we cannot know internal regulatory practice. We do not claim "no jurisdiction in the world", "regulators already hold the data", "proof of patient harm", "a validated minimum standard", or "regulatory tradition drives architecture". Reform recommendations are tiered by gap type and confidence in the Discussion.

Data availability

All materials supporting this study are deposited on the Open Science Framework (OSF) at the review-stage URL `https://osf.io/t2g8y/?view_only=0c889d89c42d4b389826757477d30c9c`, with a permanent Zenodo DOI assigned on acceptance. The deposit contains: the codebook (versions v0 and v1); the four-anchor source-concordance traceability matrix; the core (12), full (20), and extended specifications; the source inventory across all five jurisdictions with retrieval dates and access modes; the pre-adjudication coding sheets for both annotators; the final adjudicated coding sheet; the adjudication log with annotator-A and annotator-B codes, final code, and decisive evidence source for every changed cell; the FDA AI keyword list and the locked bilingual NMPA UDI keyword set; the FDA AI-list / openFDA matching pipeline; the n=70 product-family raw JSON dump and rebuilder script; the NMPA UDI metadata (PROVENANCE, FIELD_DICTIONARY, unzip log, MD5 verification record); the NMPA UDI keyword-protocol pipeline and outputs (per-tier matches, product-code clustering, transparency/verification samples); the source-archive-gap resolution note (TGA / EUDAMED / openFDA); derived tables and figures; and a README with retrieval dates and reproduction instructions.

The architecture audit source archive was locked on 2026-05-01. The FDA AI-list CSV used for the identifiability stress test was downloaded separately on 2026-05-14 from `fda.gov/media/178541/download` (MD5 `eb59f1afdc02543d98b6fc7a08de0e32`, 123,524 bytes, 1,430 data rows), because the FDA AI list is the reference set for the stress test rather than one of the jurisdictional architecture-audit sources. The NMPA UDI 2026-05-01 full release was retrieved on 2026-05-14 (UDID_FULL_RELEASE_20260501.zip, 284.56 MB, MD5 `493615E48D67B3DC9D48CBC63301371F` matching the NMPA-published value) and is mirrored externally rather than uploaded to OSF directly because of its size; the deposit contains the published MD5, the locally computed MD5, the schema/field dictionary, the keyword-protocol outputs, and the deterministic processing script needed to regenerate every China-section number from the source zip. The PMA identifiability test outputs and pipeline (`regen_pma_identifiability.py`, `pma_identifiability_full_raw.csv`, `pma_identifiability_summary.md`) and the pathway-stratified summary (`fda_identifiability_stratified.md`) are deposited in full. Selected source screenshots are shared where permissible. The codebook, coding sheets, and adjudication log are shared in full; sensitive comments in the adjudication log are removed.

Code availability

The IRR computation script (`irr_compute.py`, implementing linear-weighted Cohen's κ and Gwet's AC1), the source-archival script (`archive_sources.py`, headless-browser capture of regulatory pages), the FDA AI-list / openFDA matching pipelines for the 510(k) and PMA pathways (including `regen_pma_identifiability` the n=70 product-family raw rebuilder (`build_product_family_n70.py`), the NMPA UDI keyword-protocol pipeline (`run_keyword_protocol.py`, deterministic lxml streaming parse over the 1,136 XML parts), the results-regeneration scripts (paper headline numbers, FDA identifiability summary, FDA pathway-stratified summary, structured-only sensitivity), and the figure-regeneration scripts are deposited at the same DOI as the data (see Data Availability). Scripts are released under the MIT licence; codebook, coding sheets and analysis outputs are released under CC-BY 4.0.

Ethics approval

Ethics approval was not required because the study used only publicly accessible regulatory records and did not involve human participants, patient-level data, identifiable health data, or animal research.

Author contributions

Y.H. and A.C. conceptualised the study. Y.H. led data collection, source archiving, and primary analysis. Y.H. and A.C. developed the codebook, independently coded the audit cells, adjudicated disagreements, interpreted results, and wrote and revised the manuscript. Both authors approved the final manuscript.

Competing interests

The authors declare no competing interests.

Funding

No specific funding was received for this work.

Acknowledgements

The authors thank the maintainers of openFDA and the Internet Archive's Wayback Machine for the programmatic access that made the source-archive and stress-test pipelines reproducible.

References

U.S. Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Final Guidance. FDA Docket No. FDA-2022-D-2628, August 2025.

U.S. Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices (list; titled "Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices" as accessed). `https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices`, 2025. Accessed April 2026; page relocated to this URL by a pre-submission re-check (2026-06-16).

European Commission. The EUDAMED four first modules will be mandatory to use as from 28 May 2026. https://health.ec.europa.eu/latest-updates/eudamed-four-first-modules-will-be-mandatory-use-28-may-2026-2025-11-27_en, November 2025.

International Medical Device Regulators Forum. Machine Learning-enabled Medical Devices: Key Terms and Definitions. IMDRF/AIMD WG/N67, 6 May 2022.

International Medical Device Regulators Forum. Software as a Medical Device (SaMD): Key Definitions. IMDRF/SaMD WG/N10, 9 December 2013.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*, L 2024/1689, 2024.

Ong, A.Y., Kale, A.U., Antoun, J., Hogg, H.D.J., Hammond, B., Keane, P.A., Pearson, R., Harvey, H., and Denniston, A.K. Current challenges and the way forwards for regulatory databases of artificial intelligence as a medical device. *npj Digital Medicine*, 9:235, 2026. doi: 10.1038/s41746-026-02407-w.

Liu, Y., Yu, W., and Dillon, T. Regulatory responses and approval status of artificial intelligence medical devices with a focus on China. *npj Digital Medicine*, 7:255, 2024. doi: 10.1038/s41746-024-01254-x.

Muehlematter, U.J., Daniore, P., Vokinger, K.N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digital Health*, 3(3):e195–e203, 2021. doi: 10.1016/S2589-7500(20)30292-2.

Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019. doi: 10.1038/s41591-018-0300-7.

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E.J. AI in health and medicine. *Nature Medicine*, 28(1):31–38, 2022. doi: 10.1038/s41591-021-01614-0.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9):1337–1340, 2019. doi: 10.1038/s41591-019-0548-6.

Sendak, M.P., Gao, M., Brajer, N., and Balu, S. Presenting machine learning model information to clinical end users with model facts labels. *npj Digital Medicine*, 3:41, 2020. doi: 10.1038/s41746-020-0253-3.

Char, D.S., Shah, N.H., and Magnus, D. Implementing machine learning in health care: addressing ethical challenges. *New England Journal of Medicine*, 378(11):981–983, 2018. doi: 10.1056/NEJMp1714229.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3:199–217, 2021. doi: 10.1038/s42256-021-00307-0.

Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I.S., and Saria, S. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021. doi: 10.1056/NEJMc2104626.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., and Ghassemi, M. 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.

Babic, B., Gerke, S., Evgeniou, T., and Cohen, I.G. Algorithms on regulatory lockdown in medicine. *Science*, 366(6470):1202–1204, 2019. doi: 10.1126/science.aay9547.

Vokinger, K.N., Feuerriegel, S., and Kesselheim, A.S. Continual learning in medical devices: FDA's action plan and beyond. *Lancet Digital Health*, 3(6):e337–e338, 2021. doi: 10.1016/S2589-7500(21)00076-5.

Hwang, T.J., Kesselheim, A.S., and Vokinger, K.N. Lifecycle regulation of artificial intelligence- and machine learning-based software devices in medicine. *JAMA*, 322(23):2285–2286, 2019. doi: 10.1001/jama.2019.16842.

Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., 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.

Collins, G.S., Moons, K.G.M., Dhiman, P., Riley, R.D., Beam, A.L., Van Calster, B., 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.

Gilbert, S., Anderson, S., Daumer, M., Li, P., Melvin, T., and Williams, R. Learning from experience and finding the right balance in the governance of artificial intelligence and digital health technologies. *Journal of Medical Internet Research*, 25:e43682, 2023. doi: 10.2196/43682.

Hood, C. and Heald, D. (eds). *Transparency: The Key to Better Governance?* Oxford University Press, 2006.

Heald, D. Varieties of transparency. In Hood, C. and Heald, D. (eds), *Transparency: The Key to Better Governance?*, pp. 25–43. Oxford University Press, 2006.

Meijer, A. Understanding the complex dynamics of transparency. *Public Administration Review*, 73(3):429–439, 2013.

Star, S.L. and Ruhleder, K. Steps toward an ecology of infrastructure: design and access for large information spaces. *Information Systems Research*, 7(1):111–134, 1996.

Bowker, G.C. and Star, S.L. *Sorting Things Out: Classification and Its Consequences*. MIT Press, 1999.

Supplementary Files

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

- 2SUPPLEMENTupload.pdf