MedCrave
Step into the World of Research

International Clinical Pathology Journal

Review Article

Open Access

CrossMark

Regulatory Pathway and Evidence Requirements for Clinical Artificial Intelligence as Medical Device Software (MDSW/SAMD) Under the MDR And the AI Act.

Abstract

The aim of this review is not to catalogue every application of clinical AI, but to summarise the regulatory pathway linking intended purpose, qualification and classification, evidence requirements, governance, risk management and post-market surveillance under the European framework. When can a clinically promising algorithm be defended not only as software that performs well in development, but as software that can sustain a medical claim, survive conformity assessment and remain governable after deployment? That is where the regulatory logic of the Medical Device Regulation (MDR), the In Vitro Diagnostic Regulation (IVDR) and the AI Act converge. This review therefore treats evidence, governance and post-market surveillance not as separate downstream chapters, but as parts of the same pathway that begins with intended purpose and ends only when use in routine care remains controllable over time.

Keywords: artificial intelligence; medical device regulation; software as a medical device; clinical validation; risk management; post-market surveillance

Volume 11 Issue 1 - 2026

Marco Viscovo,[1] Marcella Marletta,[2] Roberto Verna[1,2,3]

[1]Research Center for Health and Social Wellness, Italy
[2]Academy for Health and Clinical Research, Italy
[3]World Association of Societies of Pathology and Laboratory Medicine, Italy

Correspondence: Roberto Verna, Research Center for Health and Social Wellness, University Guglielmo Marconi, Italy, Tel +3711547775, Email r.verna@unimarconi.it

Received: July 14, 2026 | **Published:** August 7, 2026

Introduction

Artificial intelligence (AI) is no longer just a generic digital technology in healthcare. Once software declares a medical intended purpose, it enters the regulatory perimeter of medical devices or in vitro diagnostic (IVD) devices and must demonstrate safety and performance in the declared context of use.[1–4] This shift matters because clinical enthusiasm and commercial adoption have often moved faster than the production of robust, transferable evidence, particularly when algorithms are validated retrospectively or on selected datasets and then expected to perform reliably in routine care.[5–7] Reporting frameworks such as CONSORT-AI, SPIRIT-AI and TRIPOD+AI were introduced precisely to make these studies more assessable,[8–10] while cybersecurity guidance and the AI Act make clear that software lifecycle control, logging, transparency and human oversight are not optional appendices but part of the compliance architecture.[11,12]

A practical consequence follows from this regulatory starting point. The same software logic can lead to very different obligations depending on the claim that is being made, the clinical role of the output, and the degree to which users are expected to rely on it. A tool framed as simple storage or neutral display is assessed very differently from software that estimates risk, supports triage, or materially contributes to diagnosis or treatment. For this reason, the path from claim to intended purpose is not a drafting exercise at the end of development, but the point at which the future conformity burden is largely determined. Once the claim becomes more clinically consequential, expectations rise accordingly: classification becomes more demanding, the role of the notified body becomes more relevant, and the evidence package has to support not only technical performance but also the plausibility that the output remains safe and useful in the declared setting of use.[2–4,13]

The aim of this review is therefore not to catalogue every application of clinical AI, but to summarize the regulatory pathway linking intended purpose, qualification and classification, evidence requirements, governance, risk management and post-market surveillance under the European framework. In practical terms, the review asks a narrower but more useful question: when can a clinically promising algorithm be defended not only as software that performs well in development, but as software that can sustain a medical claim, survive conformity assessment and remain governable after deployment. That is where the regulatory logic of the Medical Device Regulation (MDR), the In Vitro Diagnostic Regulation (IVDR) and the AI Act converge. The review therefore treats evidence, governance and post-market surveillance not as separate downstream chapters, but as parts of the same pathway that begins with intended purpose and ends only when use in routine care remains controllable over time. The starting point is straightforward: favourable internal performance is never enough on its own; what matters is whether performance and safety remain defensible in the declared context of use and over time.[2–4,7,11,12]

Qualification, classification, and evidence requirements

From a practical regulatory perspective, the key point is not simply whether an algorithm performs a medical task, but whether the manufacturer defines that task in a way that is specific enough to be assessed. A weak or generic intended purpose may seem convenient at an early development stage, yet it usually creates downstream problems: the applicable classification rule becomes harder to justify, the technical file becomes less coherent, and the evidence package risks drifting away from the actual claim. This is one reason why qualification, classification and evidence planning should be treated as one sequence rather than as separate exercises.[2–4,13] Selected IVDR and companion-diagnostic pathways are therefore kept in the review only where they affect the Medical Device Software (MDSW)/ Software as a Medical Device (SaMD) regulatory route.[3,4,14]

When evidence is considered in this light, external validation becomes more than a methodological preference. It is part of what makes a claim transferable beyond the development dataset. Calibration, handling of missingness, subgroup behaviour and leakage control all matter because regulators and users are not interested only in headline performance, but in whether the model behaves plausibly in the population and workflow for which it is proposed. This is also why reporting frameworks and appraisal tools are useful in a device context: they help distinguish studies that are merely promising from studies that are actually defensible when the claim is tied to patient management or diagnostic decision-making.[10,15–18]

What makes this particularly important for AI is that strong-looking model performance can still be regulatory weak evidence when the route from development to use is poorly described. External validation matters because it tests whether results survive movement across centers, populations, prevalences, and operational pathways rather than remaining tied to one dataset.[19,20] Calibration matters because a model that discriminates reasonably well may still misstate risk in a clinically relevant way. Leakage, outcome ambiguity, and absent subgroup analyses matter because they create an illusion of robustness that may disappear in deployment. In practice, therefore, evidence requirements are not satisfied by a high AUROC alone: they depend on whether the design, reporting, and appraisal framework make the claim transparent enough to be challenged, reproduced, and defended in front of assessors, users, and post-market reviewers.[10,15–18]

In the European framework, software qualification depends on intended purpose rather than platform. Medical Device Coordination Group (MDCG) guidance and the International Medical Device Regulators Forum (IMDRF) framework are aligned on this point: software may qualify as Medical Device Software (MDSW) whether it runs on a workstation, a smartphone, the cloud or as part of a larger device, provided that it creates clinically relevant information used for diagnosis, prognosis, monitoring, treatment or prevention.[4,13] The regulatory boundary is therefore not defined by technical sophistication, nor by whether a model uses machine learning, but by what the software is claimed to do and by the role of its output in the clinical pathway. This is also why the distinction between neutral storage or visualization and software that generates or refines clinically actionable information remains decisive. In practical terms, wording that looks minor at product level (for example whether the tool merely displays information or supports diagnosis or triage) can completely change qualification and the evidentiary burden that follows. The same logic also covers selected companion-diagnostic and laboratory software scenarios under the IVDR and, where relevant, consultative pathways with medicinal authorities.[3,4,14]

Classification then determines the conformity pathway. Under the MDR, Rule 11 governs most decision-support and monitoring software; under the IVDR, the decisive issue is how much the software contributes to the diagnostic result and how critical the clinical scenario is.[2-4] The practical consequence is that qualification and class shape not only notified body involvement, but also the required depth of evidence, technical documentation, change control and post-market obligations. In other words, the step from qualification to class is not a formality: it determines how much independent review is expected, how robust the benefit-risk demonstration must be and how tightly changes will be scrutinized over time. A stronger clinical claim usually implies a heavier regulatory architecture, which is why intended purpose cannot be treated as a late marketing adjustment. Once development has already been shaped around a weak or ambiguous claim, it is often difficult to retrofit the broader evidence package and control structure needed for a stronger one.[2–4]

Evidence requirements should be proportionate to both the claim and the risk. For diagnostic accuracy studies, the Standards for Reporting Diagnostic Accuracy Studies (STARD) statement remains the basic reporting reference; for predictive models, transferability depends not only on discrimination but also on calibration, external validation and transparent handling of data leakage, missingness and clinically relevant subgroups.[10,15–17] PROBAST and PROBAST+AI are useful because they make recurrent weaknesses explicit, including non-independent validation, poorly defined outcomes and untested applicability.[16,18] The regulatory point is not simply whether a study is publishable, but whether it supports the exact claim the manufacturer intends to make. A triage aid, a diagnostic support tool and a rule-out strategy do not require identical evidence, even when they rely on similar modelling approaches. They also do not tolerate the same degree of uncertainty about subgroup behaviour, error distribution and operational consequences. In practice, retrospective single-centre evidence may be enough for limited and tightly bounded claims, but high-impact decision-support claims require designs that are more likely to survive regulatory scrutiny and real-world deployment, including meaningful comparators, external validation and a plausible link to real clinical use.[2,10,16–20]

Review methods and core evidence

This review focuses on evidence from PubMed from 2019 to 2025 that is relevant to AI systems in imaging and laboratory/in vitro diagnostic (IVD) settings. Priority was given to studies that included multicentre data or external validation.[10,16,19,20] The intention was not to produce an exhaustive meta-analysis, but rather to identify evidence that could potentially be used to support a regulatory discussion at device level. This reflects a regulatory concern more than a bibliographic one: once a model is transferred across sites, populations and workflows, its apparent performance may deteriorate. This is precisely the point at which attractive proof of concept is no longer sufficient for a device-level claim. The review was therefore structured to prioritise studies that demonstrate transferability, whether through multicentre recruitment, programme-based implementation, pragmatic design, or explicit external validation. The inclusion and exclusion criteria are summarised in Table 1, and the data extraction template used for the review is shown in Table 2.

Table 1 Inclusion and exclusion criteria

Element	Inclusion	Exclusion
Period	2019–2025	<2019
Domain	Imaging (mammography; computer-aided detection (CADe) colonoscopy) and selected laboratory/IVD applications	Generic electronic health record (EHR)-only AI without imaging or laboratory component
Design	RCTs, pragmatic multicentre studies, or external validation studies	Single-centre studies without external validation
Outcome	Interpretable clinical/operational metrics	Decontextualized technical metrics or evident leakage
Transferability	Clear setting, population and comparator	Undefined or highly selected population without discussion

Citation: Viscovo M, Marletta M, Verna R. Regulatory Pathway and Evidence Requirements for Clinical Artificial Intelligence as Medical Device Software (MDSW/SAMD) Under the MDR And the AI Act. *Int Clin Pathol J.* 2026;11(1):33–38. DOI: 10.15406/icpjl.2026.11.00225

Table 2 Data-extraction template

Field	Description
Reference	Author, year, journal, DOI/PMID
Domain	Imaging or selected laboratory/IVD applications
Intended use	Screening, triage, reader assistance, rule-out, autoverification
Population and setting	Age, context, multicentre yes/no
Input / output	Images, signals, test panels; score, alert, classification, triage
Comparator	Standard of care or clinically relevant alternative
Primary endpoints	Defined a priori and clinically interpretable
Performance	AUROC, sensitivity/specificity, predictive values, workload, APC/ADR, etc.
External validation	Where and how validation was performed
Declared limitations	Bias, generalisability, follow-up, drift
Implications for claim	Which claim the study may support, and with what caveats

The most mature evidence in this literature is still found in imaging, particularly mammography and colonoscopy, where multicentre and pragmatic designs are beginning to emerge.[5,6,21,22]

This maturity is relative rather than absolute. Compared with many exploratory AI papers, these studies are more robust because they are closer to programme-based screening, routine endoscopy and other operational settings, where comparator choice, workload effects and clinically interpretable endpoints are important. In the laboratory/in vitro diagnostic (IVD) domain, the evidence base is more heterogeneous. However, some multicentre studies already demonstrate how model performance can be useful when the design is closer to operational conditions. Examples include rule-out strategies based on routine blood tests and machine-learning approaches to sample misidentification detection.[23,24] However, even here, robustness depends more on setting, prevalence and workflow assumptions than on a single attractive performance number. What remains consistent across domains is that high performance in the development phase does not automatically translate into deployable evidence; once the centre, population, workflow or prevalence changes, the calibration, robustness and clinical usefulness may also shift.[19,20] This is precisely why multicentricity and external validation are not just methodological virtues here, but also part of what makes the evidence regulatorily credible(Table 3).[10,16]

Table 3 Core studies with multicentre or external validation

Domain	Study and setting	Comparator	Main result / regulatory relevance
Imaging – mammography	McKinney et al.[5]; UK and USA datasets	Human readers / baseline	Absolute reduction in false positives and false negatives; useful as evidence of cross-setting generalisation.
Imaging – mammography	MASAI safety analysis[6]; randomised non-inferiority trial, 4 Swedish sites	Standard double reading	Cancer detection remained clinically acceptable while workload decreased by about 44%.
Imaging – mammography	MASAI screening performance[21]; randomised programme-based study, 4 sites	Standard double reading	Cancer detection increased with major workload reduction; recall did not increase significantly.
Imaging – colonoscopy	COLO-DETECT[22]; pragmatic multicentre randomised trial	Standard colonoscopy	Adenoma and polyp detection improved without a clear safety penalty in routine practice.
Laboratory / triage	Plante et al.[23]; 66 hospitals, external validation	PCR / COVID status	Routine laboratory data supported a rule-out strategy with high negative predictive value.
Laboratory / quality	Seok et al.[24]; retrospective multicentre study with external validation	Conventional delta check	Machine learning outperformed conventional delta-check approaches for sample-misidentification detection.

Governance, risk management and the post-market lifecycle

In operational terms, the lifecycle of AI-based devices cannot be cleanly divided into pre-market and post-market activities. A device's real-world performance may shift even if it remains technically unchanged, due to changes in scanners, analyte distributions, coding practices, user behaviour, or local workflows. For this reason, the post-market phase is not merely a legal follow-up to certification; it is an opportunity for the manufacturer to demonstrate that the original benefit-risk assessment remains valid in real-world usage conditions.[2,12,19,20,25]

This also explains why the roles of manufacturer and deployer are complementary rather than interchangeable. The manufacturer defines the intended purpose, evidence strategy, change-control logic, and instructions for use content. The deployer determines whether this logic remains workable in a local environment where data pipelines, staff competencies, override practices, and cybersecurity arrangements may differ significantly from those assumed during development. In a mature governance model, these two levels are connected by documentation, feedback, logging, and escalation processes, rather than being treated as isolated responsibilities.[2,11,12,25]

This distinction is important because clinical AI can fail not only due to a flawed algorithm, but also due to uncertainty regarding the roles of those responsible for local validation, managing overrides, reviewing complaints and escalating performance signals into regulatory or clinical actions. While the manufacturer cannot offload human oversight to the deployer through a generic instruction for use, the deployer also cannot assume that CE marking removes the need for local governance. The relevant question in practical terms is always the same: who decides, who can override, who documents and who acts when performance, safety or compliance signals change?

The risk management of AI-based devices differs from that of traditional deterministic software because hazards can arise

simultaneously from data, model behaviour, human interaction and cybersecurity. Dataset representativeness, data lineage, and subgroup analyses are important because bias and drift can cause performance to become unstable;[12,19] calibration and external validation are important because a technically robust model may still fail when moved to a new operational context;[10,16,19] usability and human oversight are important because automation bias and de facto off-label use can create risk even when the output is formally correct;[12,13] and cybersecurity is important because a security incident can result in a clinical incident through loss of integrity, availability, or control.[2,11,12] This is why the risk file cannot be limited to generic software hazards. In AI-based products, sources of harm lie partly in the code, partly in the data, and partly in the interaction between the tool and its users. A plausible risk-management strategy must therefore connect dataset controls, validation strategy, usability assumptions, override logic and security safeguards, rather than treating them as separate entities.

The practical question is not whether each domain has been mentioned somewhere, but whether they are linked in such a way that corrective action can be taken when the device is actually used.[2.10–13] Change management is therefore central. European practice still favours either a locked model updated through controlled releases or a formally bounded adaptive strategy with documented evaluation, versioning and traceability.[2,4,12] It is important to recognise that an update cannot be treated as a neutral software event, as it may alter calibration, subgroup behaviour, compatibility with local infrastructures, or the interpretation of outputs. Internationally, the U.S. Food and Drug Administration (FDA) has moved in the same direction with the proposed total product lifecycle approach, the discussion paper on AI/ML SaMD modifications and more recent guidance on predetermined change control plans and real-world evidence.[26–28] Post-market monitoring for AI devices therefore cannot remain limited to incident reporting; it must also include defined performance metrics, drift indicators and escalation thresholds, as well as sources of real-world data, such as electronic health record (EHR) systems, picture archiving and communication systems/radiology information systems (PACS/RIS), laboratory information systems (LIS) platforms, device logs, complaints and audits. Where necessary, it should also include post-market clinical follow-up (PMCF) studies.[2,12,25,28,29] These sources are important because they enable the manufacturer and deployer to observe not only whether the system fails, but also how it behaves as input quality, prevalence, workflow, and software versions evolve. In practice, PMS becomes the place where change control, performance verification and operational accountability converge.[2,25,29] Privacy is part of the design process, not an afterthought added at the end. Data protection by design and by default, together with proportional security controls, determine whether logging, retraining, monitoring and access management can be justified in practice.[30, 31] This is particularly important in healthcare, where the same architecture often has to support model monitoring, auditability and user accountability without every data flow becoming an uncontrolled reuse channel. Questions such as which data are retained, which logs contain personal information, how pseudonymisation is handled, and the basis on which monitoring data can be repurposed, are therefore not external to device governance. Rather, they form part of the criteria by which the entire lifecycle is assessed for legal and operational sustainability. In short, a compliant AI-based device is not simply one that works on a dataset; it is one that can be governed as claims, data, users, software versions and operating conditions evolve.[2,3,12,30]

Discussion and conclusions

The practical value of this review lies in linking these strands together. The literature on reporting quality and dataset shift is not separate from regulation; it explains why certain claims are difficult to sustain once the product leaves the development setting.[7,10,16,19,20] In the same way, the legal emphasis on post-market surveillance, documentation and oversight is not bureaucratic excess, but a response to the fact that software-based performance can degrade, drift or be misapplied even when the product initially appears strong.[2,25] Read together, the methodological and regulatory literature support the same conclusion: reliability has to be demonstrated as a maintained property, not as a one-time snapshot.[2,7,25]

This is also the point at which the continuity with the broader thesis and presentation material remains most visible. The central message is still that it is not enough for clinical AI to work under selected conditions; it must also enter the regulatory perimeter on defensible terms, be supported by evidence that matches the claim, and remain governable over time through risk management, change control and post-market feedback. That combination of claim discipline, evidence discipline and lifecycle discipline is what ultimately turns software from a promising tool into a medical product that can be justified in practice.[2–4,11,12,25]

Two broader implications deserve emphasis. One is methodological: if evidence is expected to support a device claim rather than a promising prototype, then reporting discipline, external validation, calibration, and operational context cannot be treated as optional refinements.[10,16,18–20] The other is organisational: compliance for clinical AI is sustained less by a single pre-market milestone than by the quality of lifecycle governance afterwards.[2,12,25–31] That is why the most persuasive dossiers are usually those in which claim, classification, evidence generation, change control, human oversight, and post-market monitoring have been built as one coherent system rather than as separate regulatory chapters assembled late in development.[2,10,12,25]

Taken together, the material reviewed here points to three conclusions. First, intended purpose is the true entry point of regulation: it determines whether software is in scope, how it is classified and what evidence will be expected.[2–4,13] This is why apparently similar tools can end up in very different regulatory positions once their declared role in diagnosis, monitoring or decision support becomes clear. Second, evidence quality remains the main bottleneck for translation into a compliant product. The literature contains many promising results, but lack of external validation, incomplete reporting, limited calibration analysis and poor operational transferability still weaken the defensibility of ambitious claims.[7,10,16,18–20] Third, post-market governance is becoming the decisive test of maturity for clinical AI, because software can only remain compliant if performance, safety, changes and cybersecurity are controlled as part of one lifecycle system.[2,11,12,25–31] The argument is therefore cumulative: claim, evidence and lifecycle governance are not separate chapters, but successive parts of the same regulatory logic.[2,11,12,25]

These conclusions also clarify the practical challenges ahead. For manufacturers, the priority is to align claim, class, evidence plan and lifecycle governance from the beginning, including recognised methods for analytical performance where laboratory or IVD claims are being made.[3,9,10,32] For healthcare institutions, the central issue is local validation and operational oversight, because population, instrumentation, data pipelines and workflow differences can all shift

Citation: Viscovo M, Marletta M, Verna R. Regulatory Pathway and Evidence Requirements for Clinical Artificial Intelligence as Medical Device Software (MDSW/SAMD) Under the MDR And the AI Act. *Int Clin Pathol J.* 2026;11(1):33–38. DOI: 10.15406/icpjl.2026.11.00225

the risk profile after deployment.[19,20,33] For regulators and notified bodies, the priority is to value evidence with real transferability[7,9,10] while minimising avoidable variability in assessment and transitional burden, especially while the European Database on Medical Devices (EUDAMED) remains only partially operational.[34]

The present review has limitations: it is narrative rather than quantitative, the underlying literature is heterogeneous, and the focus is intentionally European.[7,9,10,12] It also compresses a broad regulatory and methodological field into a short review format, which necessarily leaves some implementation details and sector-specific issues in the background. Even so, the direction is clear. Clinical AI will not be judged only by statistical performance, but by whether it can function as a medical device that remains safe, traceable and defensible in the real world. That, in the end, is the common thread linking the MDR, the IVDR and the AI Act.[2,3,12]

Acknowledgments

None.

Conflicts of interest

The authors declare that they have no potential conflicts of interest to disclose.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56.
2. European Union. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Directives 90/385/EEC and 93/42/EEC. *Off J Eur Union*. 2017;L117:1–175.
3. European Parliament and Council of the European Union. Regulation (EU) 2017/746 of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU. *Off J Eur Union*. 2017;L117:176–332.
4. Medical Device Coordination Group. *Qualification and Classification of Software: Regulation (EU) 2017/745 and Regulation (EU) 2017/746*. MDCG 2019–11 Rev. 1. 2025.
5. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89–94.
6. Lång K, Josefsson V, Larsson AM, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol*. 2023;24(8):936–944.
7. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020;368:m689.
8. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364–1374.
9. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351–1363.
10. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
11. Medical Device Coordination Group. *Guidance on Cybersecurity for Medical Devices*. MDCG 2019-16 Rev. 1. 2020.
12. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *Off J Eur Union*. 2024;L:2024/1689.
13. International medical device regulators forum. *Software as a medical device: possible framework for risk categorization and corresponding considerations*. IMDRF/SaMD WG/N12FINAL:2014.
14. European medicines agency. *Q&A for applicants, marketing authorisation holders of medicinal products and notified bodies regarding medicines used in combination with medical devices and consultation procedures for certain medical devices*. EMA/37991/2019 Rev. 6. 2025.
15. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527.
16. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51–58.
17. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008.
18. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505.
19. Guo LL, Pfohl SR, Fries J, et al. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Appl Clin Inform*. 2021;12(4):808–815.
20. Dos Santos Silva GF, Filho FNB, et al. Strategies for detecting and mitigating dataset shift in machine learning for health predictions: a systematic review. *J Biomed Inform*. 2025;170:104902.
21. Hernström V, Josefsson V, Sartor H, et al. Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study. *Lancet Digit Health*. 2025;7(3):e175–e183.
22. Seager A, Sharp L, Neilson LJ, et al. Polyp detection with colonoscopy assisted by the GI Genius artificial intelligence endoscopy module compared with standard colonoscopy in routine colonoscopy practice (COLO-DETECT): a multicentre, open-label, parallel-arm, pragmatic randomised controlled trial. *Lancet Gastroenterol Hepatol*. 2024;9(10):911–923.
23. Plante TB, Blau AM, Berg AN, et al. Development and external validation of a machine learning tool to rule out COVID-19 among adults in the emergency department using routine blood tests: a large, multicenter, real-world study. *J Med Internet Res*. 2020;22(12):e24048.
24. Seok HS, Yu S, Shin KH, et al. Machine learning-based sample misidentification error detection in clinical laboratory tests: a retrospective multicenter study. *Clin Chem*. 2024;70(10):1256–1267.
25. Medical device coordination group. *Guidance on periodic safety update report (PSUR) According to regulation (EU) 2017/745 (MDR)*. MDCG 2022–21.
26. US food and drug administration. *Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: guidance for industry and food and drug administration staff*. Issued 18, 2025.
27. US food and drug administration. *Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based Software as a Medical Device (SaMD): Discussion paper and request for feedback*. 2019.

28. US food and drug administration. *Use of real-world evidence to support regulatory decision-making for medical devices: guidance for industry and food and drug administration staff*. Final guidance. 2025.

29. Cuocolo R, Bernardini D, Pinto dos Santos D, et al. AI medical device post-market surveillance regulations: consensus recommendations by the European Society of Radiology. *Insights Imaging*. 2025;16(1):275.

30. European Union. Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation). *Off J Eur Union*. 2016;L119:1–88.

31. European data protection board. *Guidelines 4/2019 on article 25 data protection by design and by default*. Version 2.0. Adopted, 2020.

32. Clinical and laboratory standards institute. *EP17-A2: Evaluation of detection capability for clinical laboratory measurement procedures; Approved guideline—second edition*. CLSI; 2012.

33. Subasri V, Krishnan A, Kore A, et al. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. *JAMA Netw Open*. 2025;8(6):e2513685.

34. Medical device coordination group. *Guidance on harmonised administrative practices and alternative technical solutions until Eudamed is fully functional (for regulation (EU) 2017/746 on In Vitro diagnostic medical devices)*. MDCG 2022–12.

Citation: Viscovo M, Marletta M, Verna R. Regulatory Pathway and Evidence Requirements for Clinical Artificial Intelligence as Medical Device Software (MDSW/SAMD) Under the MDR And the AI Act. *Int Clin Pathol J*. 2026;11(1):33–38. DOI: 10.15406/icpjl.2026.11.00225