AEQUITAS

Regulating Artificial Intelligence in Public Healthcare to Prevent Gender and Racial Bias

A Charter-Aligned Regulatory Model for Biomedical AI

Policy document

Co-funded by
the European Union

Authors:
Alexander Gerganov, Director, Sociological Program, Center for the Study of Democracy
Victoria Bogdanova, Senior Analyst, Sociological Program, Center for the Study of Democracy

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them. Project code: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Introduction

Artificial intelligence (AI) is increasingly integrated into healthcare systems across Europe, supporting diagnosis, clinical decision-making, treatment planning, and patient risk prediction. These technologies offer considerable opportunities for improving healthcare efficiency and outcomes. At the same time, growing evidence shows that **AI systems can reproduce or amplify existing inequalities in healthcare**.[1]

Several mechanisms can lead to biased outcomes in AI-supported decision-making in healthcare. These include imbalances in training data, measurement errors in clinical variables, inappropriate proxy indicators, and differences between development environments and real-world deployment settings.[2] Bias can also arise when AI systems are introduced into clinical workflows without adequate training for healthcare professionals or without substantial human oversight and accountability mechanisms to monitor their performance across diverse patient populations.[3]

Furthermore, empirical research has demonstrated that such **biases can have serious consequences in clinical contexts**. Studies have reported persistent disparities in the diagnosis and treatment of major conditions. Women and racial/ethnic minority patients with cardiovascular disease often receive delayed or less intensive diagnosis and management compared with men and White patients,[4] and racial and ethnic differences affect diabetes care and outcomes,[5] while mental health conditions such as depression are under-diagnosed and under-treated in both women and minority populations.[6] A well-documented example is an exploratory study in the *Journal of Medical Internet Research* which found sex-based performance disparities in machine-learning algorithms used for cardiac disease prediction, in particular that female patients were consistently underrepresented in training datasets and that models showed higher false-negative rates for women, demonstrating how biased data can contribute to **inequitable diagnostic performance**.[7] Extending this concern to racial equity, another study has found that AI models trained to predict depression severity from social media language performed well for White participants but poorly for Black participants,

---

[1] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y., *Artificial intelligence in healthcare: past, present and future*, *Stroke and Vascular Neurology* 2(4): 230–243, 2017.

[2] Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S., *Mitigating bias in machine learning for medicine*, *Communications Medicine* 1, 25, 2021.

[3] Panesar, S., Kliot, M., Parrish, R., Fernandez-Miranda, J., Cagle, Y., & Britz, G.W., *Promises and Perils of Artificial Intelligence in Neurosurgery, Neurosurgery* 87(1): 33–44, 2020.

[4] Balla, S., Gomez, S.E., & Rodriguez, F., *Disparities in Cardiovascular Care and Outcomes for Women From Racial/Ethnic Minority Backgrounds*, *Current Treatment Options in Cardiovascular Medicine* 22(12):75, 2020.

[5] Young, C. & Myers, A. K., *Racial and Ethnic Disparities in Diabetes Clinical Care and Management: A Narrative Review*, *Endocrine Practice* 29(4): 295–300, 2023.

[6] Sorkin, D.H., Ngo- Metzger, Q., Billimek, J., August, K.J., Greenfield, S., & Kaplan, S.H., Underdiagnosed and Undertreated Depression Among Racially/Ethnically Diverse Patients With Type 2 Diabetes, *Diabetes Care* 34(3): 598–600, 2011.

[7] Straw, I., Rees, G., & Nachev, P., *Sex- Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: Exploratory Study*, *Journal of Medical Internet Research* 26: e46936, 2024.

revealing race-based performance disparities that reflect differences in expression patterns not captured by the models.[8]

Addressing these bias risks requires **regulatory models that translate international and European ethical and legal principles into concrete governance mechanisms** that healthcare institutions can implement in practice. Within the European Union (EU), the regulatory landscape has evolved rapidly with the adoption of the Artificial Intelligence Act (AI Act), which establishes obligations for high-risk AI systems and deployers such as hospitals.[9] However, applying these legal requirements as practical procedures within healthcare institutions remains a significant challenge.

This policy document proposes **a regulatory model for biomedical AI governance which aligns with the EU acquis**, including the EU Fundamental Rights Charter, and is designed for public healthcare systems and civil society organisations (CSOs). The proposed model is anchored in European regulatory instruments, including the AI Act, the Medical Device Regulation (MDR), the In Vitro Diagnostic Regulation (IVDR), the General Data Protection Regulation (GDPR), and the emerging European Health Data Space (EHDS). It translates fundamental rights principles, particularly human dignity, equality, non-discrimination, and the right to healthcare, into practical safeguards applicable to the full lifecycle of AI systems. References to selected examples from non-EU jurisdictions are included only where they provide transferable insights and are intended to complement the European legal and rights-based framework underpinning the model.

Further, the regulatory model draws on established regulatory approaches from other high-risk sectors, including aviation safety management, financial model risk governance, pharmacovigilance in medicines regulation, and algorithmic impact assessments in public administration. These approaches provide valuable lessons for building governance frameworks capable of addressing the complex lifecycle risks associated with biomedical AI and have been adapted here to explicitly embed gender and racial equity considerations throughout the AI lifecycle.

The model provides healthcare institutions and CSOs with a **practical governance framework covering procurement, validation, deployment, monitoring and accountability** for biomedical AI systems. By embedding fundamental rights protections into operational processes, this framework aims to reduce risks of gender and racial bias in AI-supported healthcare decision-making and contribute to more equitable healthcare outcomes across the EU. In addition, the document **offers a set of practical policies, tools and artefacts that hospitals and CSOs can adapt to different national contexts** to implement the model consistently in real-world settings.

---

[8] Rai, S., Stade, E. C., Giorgi, S., Francisco, A., Ungar, L. H., Curtis, B., & Guntuku, S. C., *Key language markers of depression on social media depend on race*, *Proceedings of the National Academy of Sciences* 121(14): e2319837121, 2024.

[9] European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council, Laying Down Harmonised Rules on Artificial Intelligence (AI Act)*, PE/24/2024/REV/1, *Official Journal of the European Union*, 13 June 2024.

The European Legal Framework for AI in Healthcare

The EU has established a multi-layered regulatory environment for biomedical AI, consisting of four complementary legal layers: the AI Act, medical device law and related guidance, data protection, and health-data governance.

The AI Act establishes a risk-based regulatory framework for AI systems. Many healthcare applications, including clinical decision-support tools and diagnostic algorithms, fall into the category of high-risk AI systems, which are to comply with strict requirements relating to risk management, data governance, transparency, human oversight, and accuracy. Applicable requirements also include the identification and mitigation of risks of discriminatory outcomes through disaggregated analysis of populations, treating sex and gender, ethnic origin and other social categories as cross-cutting variables. **Healthcare institutions that deploy AI systems also have specific obligations under the AI Act**. These include ensuring that systems are used in accordance with the provider's instructions, monitoring their performance, maintaining operational logs, and reporting serious incidents to relevant authorities, inter alia where such incidents may disproportionately affect different patient populations.

In terms of the AI Act's application, in 2025, prohibited practices already applied and some oversight and governance rules also started.[10] The next major milestone is August 2026, when the majority of obligations under the Act become applicable (including Annex III high-risk systems and Art. 50 transparency duties). High-risk AI embedded in regulated products (which includes many medical device AI systems) is scheduled to apply from August 2027.

For hospitals, the most practically important AI Act provisions are:

- **High-risk system requirements**. Providers are to implement a lifecycle risk management system (Art. 9) and data governance practices with: explicit bias detection and mitigation expectations (Art. 10), including assessment of data representativeness and potential performance disparities across populations using disaggregated analysis by sex and gender and racial or ethnic origin; transparency safeguards and provision of information to deployers (Art. 13); human oversight designed to prevent automation bias and enable override or stop procedures (Art. 14); and accuracy, robustness and cybersecurity controls with explicit attention to feedback loops and AI-specific attacks (Art. 15), which may otherwise exacerbate existing inequities if left unaddressed.

- **Deployer obligations**. Hospitals and other deployers have to use high-risk AI systems in line with provider instructions (Art. 26); assign competent human oversight (Art. 26); ensure input data are representative and complete where the deployer controls data inputs (Art. 26), also (where feasible) consideration of demographic diversity in clinical populations; monitor system operation (Art. 26); retain automatically generated logs for at least six months (Art. 26); and timely report serious incidents to providers and competent authorities (Arts. 26, 73), particularly where there is a risk of unequal clinical outcomes.

[10] European Commission, *Timeline for Implementation of the EU AI Act*, Brussels: European Commission, 2024.

- **Fundamental Rights Impact Assessment (FRIA)**. Before deploying certain high-risk AI systems, bodies governed by public law (and private entities providing public services) have to perform a FRIA, describing processes, affected groups, risks, human oversight measures, and mitigation or complaint mechanisms, and then notify the market surveillance authority (Art. 28). This requirement is directly relevant to public hospitals in many Member States, as it provides a mechanism to identify and address, prior to deployment, risks of discrimination and inequitable impacts assessed through disaggregated analysis by sex and gender and race/ethnicity.

Beyond the AI Act, **AI systems used in medical contexts are also subject to medical device law and related guidance**, including the *Medical Device Regulation (MDR)*[11] and the *In Vitro Diagnostic Regulation (IVDR),*[12] which regulate the safety and performance of medical technologies. These regulations require clinical evaluation and post-market surveillance for medical device software, including AI-based tools, with attention to equitable performance across disaggregated patient populations where data allow. Further, The Medical Device Coordination Group's (MDCG) *Guidance on Qualification and Classification of Software under MDR/IVDR* provides guidelines on how to qualify and classify software under the MDR and IVDR.[13] It emphasises that classification depends on the intended purpose rather than the software's location (e.g., cloud-based or embedded), and includes examples drawn from hospital information systems and clinical decision-support software, highlighting the need to consider differential impacts on diverse patient populations. The MDCG's *Guidance on Clinical and Performance Evaluation of Medical Device Software* provides additional support, reinforcing the expectation that evidence should be proportionate to the software's clinical role, including evidence that AI performance is consistent across populations differentiated by sex and gender and racial or ethnic background.[14]

The relationship between the AI Act and existing medical device law is clarified in the *Guidance on the Interplay between the AI Act and the MDR/IVDR*.[15] The document explains that the AI Act applies alongside these two regulations. In practice, this means many **medical AI systems have to comply with both frameworks simultaneously**. The guidance also clarifies when medical AI qualifies as high-risk under the AI Act, which is typically the case when the AI system itself is a medical device (or a safety component of one) and therefore subject to third-party conformity assessment by a notified body. For hospital procurement, this has practical implications. Hospitals should require vendors to demonstrate the regulatory status of their systems, including MDR/IVDR classification, involvement of a notified body where applicable, and alignment with the AI Act's high-risk requirements, as well as documentation

---

[11] European Union, *Regulation (EU) 2017/745 on Medical Devices (MDR)*, *Official Journal of the European Union*, L117, Brussels: EU, 2017.

[12] European Union, *Regulation (EU) 2017/746 on In Vitro Diagnostic Medical Devices (IVDR)*, Official Journal of the European Union, L117, Brussels: EU, 2017.

[13] Medical Device Coordination Group, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 (MDR) and Regulation (EU) 2017/746 (IVDR)*, Brussels: MDCG, 2019.

[14] Medical Device Coordination Group, *Guidance on Clinical Evaluation (MDR) and Performance Evaluation (IVDR) of Software in Medical Devices*, Brussels: MDCG, 2020.

[15] Medical Device Coordination Group & European AI Board, *Guidance on MDR/IVDR and AI Act Interplay*, Brussels: European Commission, 2025.

of disaggregated population group performance and bias mitigation measures where relevant. Procurement contracts should also require vendors to provide the documentation, transparency information, and oversight measures needed by deployers under the AI Act, including provisions to ensure equitable outcomes for all patients.

Several **additional EU instruments also matter for healthcare AI governance in practice**. The *General Data Protection Regulation (GDPR)* remains a core constraint for biomedical AI because health data are special-category personal data, and hospital deployments often require Data Protection Impact Assessments (DPIAs). The AI Act (Arts. 26(9), 27(4)) explicitly links deployer practice to DPIA obligations (use of provider information to support DPIA) and positions FRIA as complementary to DPIA, which can include consideration of disparate impacts across populations differentiated by sex and gender and racial/ethnic origin. Furthermore, the *European Health Data Space (EHDS)* aims to facilitate secure access to health data for healthcare delivery and research while maintaining strong data protection safeguards.[16] The EHDS is expected to play a key role in enabling more representative datasets for biomedical AI development, supporting equity-focused evaluation and reducing the risk of underrepresentation of patient populations with certain demographic characteristics.[17]

Finally, as regards cybersecurity, the European Union Agency for Cybersecurity has produced guidance on cybersecurity practices for AI. It argues that existing organisational standards (e.g., ISO security management and quality management) can be adapted to AI contexts.[18] The MDCG's *Guidance on Cybersecurity for Medical Devices* outlines how manufacturers should meet MDR and IVDR cybersecurity-related requirements through risk management and minimum cybersecurity expectations.[19] Importantly, strong cybersecurity governance underpins equitable AI deployment: protecting data integrity, access, and system reliability helps prevent unintended amplification of gender and racial/ethnic inequities.

Together, these legal instruments create a complex regulatory environment in which healthcare AI systems are to operate. The AI regulatory model for European public hospitals proposed here is designed explicitly around those four core legal 'layers' that govern health AI practice. Respectively, the model foresees mechanisms to monitor and mitigate inequitable impacts across populations differentiated by sex and gender and racial or ethnic origin throughout the AI lifecycle.

European and International Governance Frameworks Applicable to Responsible AI in Healthcare

It is important to note that the European regulatory and standardisation frameworks discussed above constitute the primary legal foundation for biomedical AI governance within the Union. In addition to the EU's regulatory environment for biomedical AI, several non-

---

[16] European Union, *Regulation on the European Health Data Space (EHDS)*, Brussels: European Union, 2024.

[17] European Commission, *European Health Data Space (EHDS)*, Brussels: European Commission, 2025.

[18] European Union Agency for Cybersecurity (ENISA), *Cybersecurity of AI and Standardisation*, Athens: ENISA, 2023.

[19] Medical Device Coordination Group, *MDCG 2019- 16 Rev. 1 – Guidance on Cybersecurity for Medical Devices*, Brussels: European Commission, 2020.

binding international instruments provide guidance for managing AI risks. Notably, the **additional governance frameworks overviewed in this section are included to demonstrate implementation techniques that can be transferred into the EU context, while remaining fully subordinate to European legal requirements and fundamental rights principles**. They are used selectively as analogues from other jurisdictions where comparable lifecycle governance mechanisms have been created in practice, particularly in areas such as model risk management, algorithmic auditing, and institutional impact assessment.

For instance, the National Institute of Standards and Technology's (NIST) *AI Risk Management Framework* (RMF) offers a comprehensive approach to managing AI risks through a lifecycle approach organised around governance, risk mapping, risk measurement and risk management.[20] The framework explicitly recognises fairness and addressing harmful bias as core dimensions of trustworthy AI. Further, as part of risk mapping and measurement, it encourages organisations to identify and assess disparate impacts across patient populations, also considering sex and gender and racial/ethnic differences. The goal is to embed trustworthiness and risk controls throughout design, deployment, and monitoring, combining technical evaluation with organisational accountability. For hospitals, the AI RMF's strength lies in its ability to be translated into concrete practices, such as **defined roles for equity oversight, methodical evaluation of disaggregated population performance, and documented mitigation strategies**, while maintaining a broader rights-based perspective.

On the management-system side, the International Organization for Standardization's *ISO/IEC 42001* sets requirements for an organisational AI management system, focusing on policies, objectives, processes, and continual improvement for responsible AI.[21] While not prescriptive on specific harms, it supports the integration of fairness and non-discrimination principles into organisational policies and risk controls, enabling institutions to address potential gender and racial biases in a coordinated way. Complementary to that, *ISO/IEC 23894* provides guidance on integrating risk management into AI development and deployment, including the identification and treatment of risks related to unintended or disproportionate impacts on different population groups.[22] For clinical AI specifically, *ISO 14971* defines risk-management principles widely used in medical device regulation, which describe systematic hazard identification, risk evaluation, risk control, and lifecycle monitoring.[23] These processes can be extended to consider inequitable performance or outcomes across disaggregated populations as part of harm identification. The described standards are valuable as they suggest **a management system structure that hospitals can adopt**. They also provide a common language through which equity-related requirements, such as disaggregated population evaluation and bias mitigation, can be specified in procurement and contracting with suppliers.

---

[20] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, Gaithersburg, MD: NIST, 2023.

[21] International Organization for Standardization, *ISO/IEC 42001:2023 – Artificial Intelligence Management System*, Geneva: ISO, 2023.

[22] International Organization for Standardization, *ISO/IEC 23894:2023 – Artificial Intelligence Risk Management*, Geneva: ISO, 2023.

[23] International Organization for Standardization, *ISO 14971:2019 – Medical Devices: Application of Risk Management to Medical Devices*, Geneva: ISO, 2019.

Ethical governance frameworks also provide important normative guidance. The EC's *Ethics Guidelines for Trustworthy AI* is a useful, though non-binding, policy instrument.[24] It defines seven core requirements for responsible AI systems, including diversity, non-discrimination and fairness, which explicitly call for the avoidance of bias and the consideration of impacts across populations with different demographic characteristics, such as sex and gender and racial/ethnic background. The *Assessment List for Trustworthy Artificial Intelligence (ALTAI)*, a voluntary self-assessment tool developed by the EC's High-Level Expert Group on Artificial Intelligence, implements those guidelines as a self-assessment list.[25] In a healthcare regulatory model, ALTAI's value lies in its **question-driven format, which can be adapted into practical procurement and deployment checklists** prompting stakeholders to consider issues such as the representativeness of training data, the risk of disparate clinical outcomes, and the adequacy of safeguards against discriminatory effects, thus embedding equity considerations into routine decision-making by clinicians, IT specialists, and hospital managers.

From a human-rights perspective, the Council of Europe's *Framework Convention on AI* is a relevant legal instrument explicitly anchored in the principles of human dignity, transparency, accountability, equality and non-discrimination, and privacy, thereby providing a legal basis for addressing discriminatory effects, including those linked to gender and racial bias in AI systems.[26] Similarly, the Organisation for Economic Co-operation and Development's *Principles on Artificial Intelligence*[27] and UNESCO's *Recommendation on the Ethics of AI*[28] offer international normative guidance that encourages the identification and mitigation of unfair or disparate impacts across disaggregated populations, reinforcing the expectation that AI systems should not reproduce or amplify structural inequalities. For health-specific ethical governance, the World Health Organization has articulated six core principles for AI in health: autonomy, well-being, explainability, responsibility, equity, and responsiveness.[29] The principle of **equity explicitly requires attention to differences in access, performance, and outcomes across diverse patient populations, also considering sex and gender and racial or ethnic differences**.

These international governance frameworks map directly to Charter-aligned requirements on dignity, equality, non-discrimination, and the right to health, and share a common emphasis on lifecycle governance – principles that can be directly translated into practical controls for healthcare institutions.

---

[24] European Commission, *Ethics Guidelines for Trustworthy AI*, Brussels: European Commission, 2019.

[25] European Commission, *Assessment List for Trustworthy Artificial Intelligence (ALTAI) – Self-Assessment*, Brussels: European Commission, 2020.

[26] Council of Europe, *Framework Convention on Artificial Intelligence (CAI)*, Strasbourg: Council of Europe, 2023.

[27] Organisation for Economic Co-operation and Development, *OECD Principles on Artificial Intelligence*, Paris: OECD, 2019.

[28] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, Paris: UNESCO, 2021.

[29] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health*, Geneva: WHO, 2021.

Lessons from Cross-Sector Regulatory Models

High-risk sectors such as aviation, banking, and pharmaceuticals have long developed sophisticated governance systems for managing technological hazards. These sectors offer valuable regulatory approaches that can be adapted to healthcare AI governance and the management of related gender and racial bias challenges.

While gender and racial bias are not technological hazards in a narrow sense, they emerge as socio-technical risks when data, algorithms, and organisational practices interact with structural inequalities in healthcare. Here, 'hazard' refers to the system-level manifestation of bias in outputs, decisions, and clinical outcomes, rather than its origin. Healthcare itself already provides extensive evidence of such effects, including documented disparities in diagnosis, treatment, and risk prediction across populations differentiated by sex and gender and race/ethnicity. Cross-sector examples therefore complement healthcare evidence by illustrating mature governance mechanisms for managing complex, high-impact systems, alongside empirical findings from healthcare that demonstrate the need for such regulatory controls.

One particularly relevant example is the *Safety Management System (SMS)* used in aviation. International civil aviation regulations treat safety not as a one-time certification exercise but as a continuous organisation-wide process enabling decisions based on hazard identification, risk management, safety assurance (monitoring and audit), management of change, and ongoing improvement.[30] Applying this logic to healthcare AI would mean **establishing an 'AI Safety Management System' that explicitly incorporates equity considerations**, including: (1) a hazard and near-miss reporting channel for clinicians capturing potential gender or racial bias incidents; (2) a formal risk log for each AI system documenting performance differences across disaggregated patient populations; (3) change-control triggers that evaluate the impact of model updates, data drift, or workflow changes on equity outcomes; and (4) periodic safety assurance reviews with corrective actions addressing both technical risks and identified gender or racial inequities. The AI Act's (Arts. 9, 72) explicit emphasis on lifecycle risk management and post-market monitoring aligns closely with this SMS logic.

A second useful approach comes from financial regulation, where complex quantitative tools are treated as 'models' requiring robust development, independent validation and governance oversight. In banking supervision, comparable risk governance frameworks have been implemented in jurisdictions such as the United States (e.g. the U.S. Federal Reserve's *Supervisory Guidance on Model Risk Management)* and provide an example of how validation, oversight, and independent model review can be implemented.[31] Similar approaches could be applied to biomedical AI systems to ensure that **clinical decision-support tools are independently validated (internally or externally) before they are introduced** into healthcare settings, with **explicit assessment of performance across populations differentiated by sex and gender, and by racial or ethnic characteristics**. Minimum validation artefacts should document the system's intended purpose, working limits, known failure

[30] International Civil Aviation Organization, *Safety Management Manual (Doc 9859)*, 4th ed., Montréal: ICAO, 2018.

[31] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, *Supervisory Guidance on Model Risk Management (SR 11- 7)*, Washington, D.C., 4 April 2011.

modes, and disaggregated population performance, including any disparities or mitigation measures applied. Revalidation should be triggered whenever the system is deployed in a new site, applied to a different patient population, connected to new equipment or protocols, updated by the provider, or shows signs of performance drift. These practices are consistent with the AI Act (Arts. 13, 14, 60), which mandates testing, transparency to deployers, and human oversight designed to detect and prevent automation bias, ensuring that AI-supported care is both safe and equitable across diverse populations.

When it comes to post-deployment governance, pharmaceutical regulation provides a strong model: in pharmacovigilance, adverse reactions and safety signals are systematically collected and analysed to detect emerging risks. In particular, the European Medicines Agency's *Good Pharmacovigilance Practices (GVP) Module VI* sets detailed expectations for collecting, managing, and submitting reports of suspected adverse reactions, reflecting an ecosystem approach to incident reporting and follow-up.[32] A comparable **'AI-vigilance' system could allow hospitals to report and analyse incidents** involving AI-supported clinical decisions, with particular attention to equity outcomes. Key elements would include: (1) systematic reporting that captures any observed inequalities in care or outcomes across sex and gender or racial/ethnic differences; (2) mandatory documentation of serious incidents where bias may have contributed to differential treatment; (3) obligations to notify the provider and relevant authorities of equity-related concerns (notifying authorities if bias leads to serious harm, such as misdiagnosis or unequal treatment, while keeping routine inequities within internal quality monitoring); and (4) trend and signal detection functions that identify emerging patterns of bias across departments, facilities, or the broader healthcare network. Such monitoring would enable proactive mitigation measures and continuous improvement, ensuring that AI-supported care is not only safe but also equitable.

Finally, public-sector governance tools such as algorithmic impact assessments and bias audits offer practical methods for evaluating the societal impact of automated decision-making systems, including potential disparities across sex and gender or racial and ethnic differences. These tools translate ethical principles and human rights standards into actionable formats that organisations can use to anticipate and mitigate inequities before deployment. For example, Canada's federal *Algorithmic Impact Assessment* scores AI systems' risks and links them to mitigation measures.[33] This tool explicitly frames deployment as an impact-driven governance problem. Further, New York City's approach to automated employment decision tools requires bias audits and disclosure obligations towards the public.[34] Respectively, hospitals and CSOs working in healthcare could implement an **'AI Impact Assessment & Audit' package that combines a structured equity-focused questionnaire with periodic independent fairness audits**, producing a publishable summary of performance across disaggregated populations and mitigation efforts. Implementing such a package would contribute to promoting transparency, public accountability, and patient trust in AI-supported care.

---

[32] European Medicines Agency, *GVP Module VI: Collection and Submission of Suspected Adverse Reactions*, London: EMA, 2017

[33] Treasury Board of Canada Secretariat, *Algorithmic Impact Assessment*, Ottawa: Government of Canada, 2026.

[34] New York City Department of Consumer and Worker Protection, *Automated Employment Decision Tools*, New York: City of New York, 2023.

The inclusion of regulatory models from sectors such as aviation safety, financial model risk management, pharmacovigilance, and algorithmic governance is intended as a deliberate transfer of established governance principles. These domains were selected because they practically implement formalised approaches to risk detection, monitoring, and mitigation in complex systems where differential impacts on individuals may arise. Several already incorporate fairness practices in working terms, including demographic risk stratification in pharmacovigilance, model validation and bias controls in financial supervision, and systematic reporting of adverse events affecting specific populations. Their relevance lies both in conceptual similarity, as well as in their proven use of lifecycle governance, mandatory reporting, and independent oversight mechanisms that can be directly adapted to address gender and racial equity risks in biomedical AI systems.

Blueprint for a Charter-aligned AI regulatory model for public hospitals and CSOs

This section sets out a regulatory model architecture which is **anchored in the EU Charter's principles, but is expressed as practical procedures that hospitals and CSOs can implement** in practice. The Charter provides the normative foundation for the governance of health technologies in Europe, particularly through the principles of human dignity, equality, non-discrimination, and the right to healthcare.[35] These rights are especially relevant when automated systems influence clinical decisions affecting patient care, including potential gaps among populations differentiated by sex and gender and racial/ethnic background.

At the most fundamental level, Art. 1 of the Charter, which establishes that human dignity is inviolable, supports governance requirements ensuring that AI tools are used in ways that preserve human judgement and respect patient autonomy. In practical terms, this principle requires **safeguards against over-reliance on automated recommendations, mechanisms for contesting AI-assisted decisions, and clinical oversight structures that prevent automation bias** - the tendency of human decision-makers to trust excessively in algorithmic outputs, which can amplify inequities if AI recommendations systematically disadvantage certain disaggregated populations.

Similarly, Arts. 20 and 21, which enshrine equality before the law and prohibit discrimination, provide the legal foundation for explicit fairness and bias-mitigation obligations in biomedical AI. Hospitals deploying AI systems should therefore implement **processes to monitor system performance across diverse patient populations**, including sex and gender, race or ethnicity, and intersectional combinations of these characteristics. Where inequities in performance or outcomes are detected, corrective actions should be implemented through model adjustment, changes in clinical workflow, or alternative decision pathways, ensuring that AI deployment does not entrench structural inequities.

What is more, the Charter's Art. 35 guarantees the right to a high level of human health protection and thus further supports the principle that AI systems may only be deployed where they demonstrably contribute to safe, effective, and equitable healthcare. AI

[35] European Union, *Charter of Fundamental Rights of the European Union*, 2012/C 326/02, Strasbourg: EU.

technologies should not introduce new barriers to care or exacerbate existing disparities affecting historically underserved or otherwise disadvantaged patient populations, and their impact on such patient populations should be evaluated continuously.

To secure these rights in practice, the regulatory model proposed in this policy document adopts a **lifecycle governance approach that structures AI oversight around a series of mandatory decision points or 'gates'**. These gates reflect the regulatory logic embedded in the EU's AI Act and related health legislation while providing hospitals with a clear and auditable governance pathway. The model assumes the hospital is usually a deployer (under the AI Act) and a healthcare provider under national health law, but it also handles the case where the hospital becomes a provider/manufacturer (as in the example of in-house AI).

Hospitals are primary deployers and implementation controllers of AI systems, while CSOs act as independent oversight, audit, and rights-based monitoring bodies without day-to-day control over clinical deployment decisions.

The implementation of this regulatory model follows a proportionality principle recognising that **hospitals and CSOs may have differing levels of technical, organisational, and analytical capacity across Member States and healthcare systems**. Mandatory minimum requirements apply uniformly to ensure baseline safety, transparency, and equity safeguards. Additionally, advanced practices, such as extended disaggregated population analysis, external auditing, and sophisticated bias monitoring tools, may be implemented progressively depending on institutional capacity. The technical annex also follows this approach by distinguishing between minimum and recommended requirements.

1. **Scope and classification gate.** Before procurement or development, institutions have to determine the regulatory status of a given AI tool. This includes assessing whether the system qualifies as medical device software under the MDR or IVDR, whether it constitutes a high-risk AI system under the AI Act, or whether it falls into a lower-risk category that still requires governance oversight. At this early stage, hospitals can flag tools whose intended use may have heightened equity implications or affect diverse patient populations, noting that these tools will require more detailed evaluation of gender and racial bias at later stages. The guidance produced by the MDCG on software qualification and classification provides practical examples that can help avoid regulatory misclassification and set the stage for bias and equity assessments.[36] Hospitals lead classification and regulatory determination. CSOs provide independent advisory input on equity-related risks but have no decision-making authority at this stage.
2. **Pre-procurement and contract gate.** Hospitals frequently exert their greatest leverage over AI governance during procurement. At this stage, institutions should ensure that suppliers commit contractually to addressing equity risks, including gender and racial bias, throughout the AI lifecycle. Suppliers should be required to provide documentation demonstrating regulatory compliance, including evidence of

[36] Medical Device Coordination Group, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 (MDR) and Regulation (EU) 2017/746 (IVDR)*, Brussels: MDCG, 2019.

conformity assessments where relevant, performance validation results, and information necessary for deployers to fulfil their obligations under the AI Act. Procurement contracts should explicitly allocate responsibilities between provider and deployer, including requirements relating to transparency, documentation, logging, human oversight, and post-deployment support. The goal at this gate is to embed bias-awareness and accountability into contractual obligations, creating a foundation for safe, equitable deployment. Hospitals are responsible for procurement decisions and contractual arrangements. CSOs may review procurement documentation for equity and rights considerations and issue non-binding recommendations.

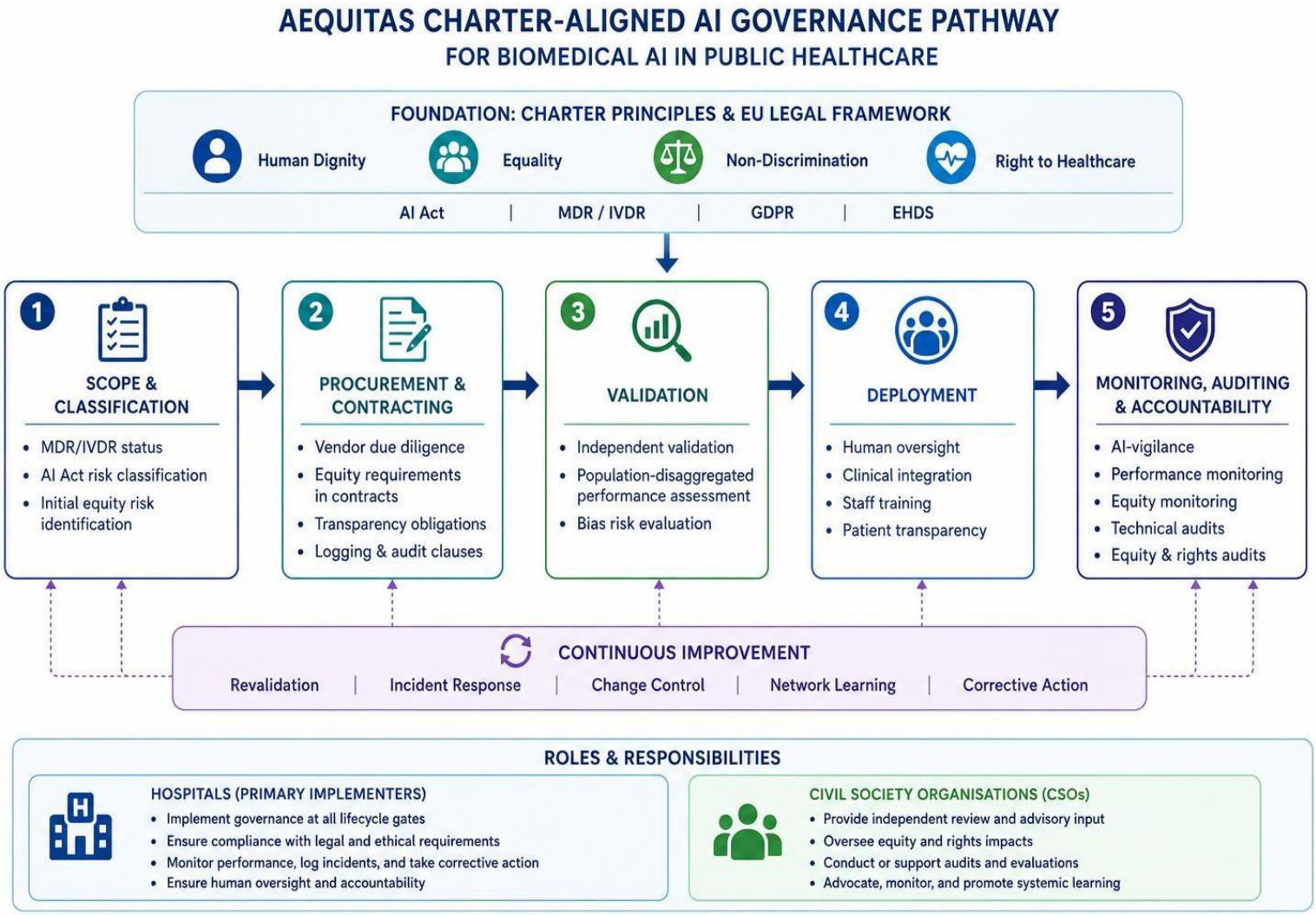
3. **Validation gate.** The third stage is a validation gate, which ensures that AI tools are independently evaluated before entering clinical workflows. Validation should assess both overall performance and outcomes across disaggregated patient populations, with particular attention to sex and gender and racial/ethnic background. This step is critical because biomedical AI systems can exhibit disparate performance across populations when training data are unrepresentative or when contextual differences between development and deployment environments are overlooked.[37] Evaluation should also document any mitigation measures applied and highlight remaining equity risks, providing hospitals with evidence to make informed deployment decisions. Hospitals are responsible for conducting or commissioning validation and interpreting results for deployment decisions. CSOs provide independent review of disaggregated performance and bias risks without control over validation outcomes.
4. **Deployment gate.** Once a system has passed validation, the fourth stage is a deployment gate. Hospitals need to ensure working conditions support safe, equitable, and responsible use. This includes assigning trained oversight personnel, implementing mechanisms that allow clinicians to override or disable algorithmic recommendations, and ensuring users understand the system's intended purpose, limitations, and any disaggregated performance considerations. Clinicians should be informed of potential unjust variations in outputs when sex and gender or racial/ethnic differences are considered, and patients should be made aware when AI-supported decision tools are being used in their care. Hospitals retain full responsibility for deployment decisions, clinical integration, and implementation oversight. CSOs have no role in deployment but may review deployment readiness documentation from an equity and rights perspective.
5. **Monitoring, auditing and accountability gate.** This final gate governs the post-deployment phase of the AI lifecycle. AI systems should be subject to continuous monitoring for overall performance, disparities across disaggregated populations, and safety or equity incidents. Hospitals should maintain logs of system operation, implement standardised incident reporting procedures that capture potential gender or racial bias, and periodically reassess system performance through revalidation exercises. Where inequities or risks of bias emerge that cannot be mitigated effectively, the system should be modified, suspended, or decommissioned. Findings should inform both internal governance and, where appropriate, broader network-level learning to prevent recurring bias. Hospitals are responsible for ongoing

[37] Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H., *Ensuring Fairness in Machine Learning to Advance Health Equity*, *Annals of Internal Medicine*, 169(12), 866-872, 2018.

monitoring, logging, incident reporting, and implementation of mitigation actions. CSOs conduct independent audits, assess systemic equity risks across sites, and support escalation of persistent disparities, without operational control over system modifications.

Together, these lifecycle gates transform fundamental rights principles into concrete governance processes capable of addressing risks of gender and racial bias in clinical AI systems.

Figure 1: AEQUITAS Charter-Aligned AI Governance Pathway



Practical policies, operational tools and governance artefacts for hospital and CSO professionals

In order to function effectively in real healthcare settings, the regulatory model proposed in this policy document foresees the creation of practical instruments that can be reused across institutions and adapted to different national contexts. It emphasises the need for standardised procedures, documentation and templates that enable both hospitals and CSOs to administer AI systems consistently. The following set of instruments aligns with the requirements of the AI Act, guidance concerning the MDR and the IVDR, and cross-sector governance practices.

Procurement and contract stage

- **Procurement and contracting package.** The EU's AI procurement community has published model contractual clauses (high-risk and non-high-risk versions) and updated them to reflect the AI Act, explicitly encouraging public authorities to use them and providing commentary for implementation.[38] For AI in healthcare, with a view to assessing and mitigating gender and racial bias risks, these clauses can be adapted into a hospital-specific annex, as a standalone artefact, that requires (at minimum):
    - (1) standardised reporting of training data composition, including sex and gender and racial/ethnic representation, to support assessment of representativeness and bias risk;
    - (2) disclosure of data provenance, including whether datasets are proprietary or non-proprietary, their accessibility conditions, and the implications for transparency, independent validation, and the assessment of gender and racial bias;
    - (3) disclosure of performance across disaggregated populations and validation evidence, including stratified results by sex and gender and racial or ethnic background, documentation of any known disparities, bias mitigation measures applied during development, and supporting validation tests (including site-specific considerations relevant to equity and fairness);
    - (4) requirements for logging and auditability, ensuring the system is capable of generating usage logs that can support independent audits for bias or disparate impacts;
    - (5) update and change-control obligations, requiring the provider to inform the hospital of model updates, retraining, or workflow changes, and to document their potential effects on performance with respect to differentiated populations;
    - (6) bias-related incident response commitments, obliging the provider to report identified errors or inequities affecting patient populations differentiated by sex and gender or racial/ethnic background and outline steps for remediation;

[38] European Commission, *Updated EU AI Model Contractual Clauses Now Available*, Brussels: European Commission, 2025.

- (7) clear allocation of responsibilities, specifying provider and deployer obligations regarding compliance with AI Act requirements and the management of equity and bias-related risks.

Validation and pre-deployment stages

- **Health-data governance and representativeness policy.** To mitigate gender and racial bias, all hospitals and CSOs deploying AI should adopt a formal policy ensuring that data used for training, fine-tuning, or implementation purposes are representative, complete, and accurate for the populations served. This includes assessing demographic composition, documenting underrepresented patient populations, and defining mitigation measures where necessary. The policy should cover both in-house and externally sourced datasets, including proprietary and open-access data, and require verification of disaggregated performance and any bias-mitigation strategies applied by providers. Such governance aligns with the AI Act (Art. 10), which requires examination and mitigation of bias risks, and mirrors best practices in the health-AI bias mitigation literature covering the full lifecycle from data collection through post-deployment monitoring.[39] This policy is not an artefact per se but a policy layer guiding other artefacts. It works as a pre-deployment principle and feeds directly into continuous monitoring, auditing, and accountability.

- **Cybersecurity baseline and AI-specific threat model.** Hospital AI systems can be vulnerable not only to conventional attacks such as data poisoning, adversarial examples, or model evasion, but also to manipulations or errors that exacerbate gender or racial bias, for example by disproportionately affecting underrepresented demographic strata in training or implementation data. The AI Act (Art. 15) requires resilience against unauthorised alterations of outputs or performance, highlighting the need for AI-specific safeguards. Hospitals should therefore establish a cybersecurity baseline and an AI-specific threat model that explicitly considers bias-related risks alongside technical threats. This living document functions as both a policy and practical tool, serving as a pre-deployment checklist to ensure implementation of cybersecurity controls aligned with the MDR and IVDR, organisational information security governance (e.g., ISO/IEC 27001), and bias-sensitive threat modelling and test plans.[40] It also supports continuous post-deployment monitoring to detect emerging technical and bias-related risks and to respond proactively. By implementing this policy, hospitals can strengthen system resilience, safeguard equity in AI-supported decision-making, and maintain compliance with the AI Act (Arts. 9-15) and MDR/IVDR requirements.

- **An 'AI system dossier' per tool.** This artefact is a standardised document maintained for each AI system prior to deployment within a healthcare organisation. The dossier should summarise the system's intended purpose, clinical workflow integration,

---

[39] Char, D.S., Shah, N.H., & Magnus, D., *Implementing Machine Learning in Health Care — Addressing Ethical Challenges*, *New England Journal of Medicine* 378: 981-983 (2018).

[40] Medical Device Coordination Group, *MDCG 2019-16 Rev. 1 - Guidance on Cybersecurity for Medical Devices*, Brussels: European Commission, 2020.

validation results, and performance metrics, with particular attention to performance across patient populations differentiated by sex and gender or racial/ethnic origin and any known disparities or mitigation measures. It should document human oversight arrangements, clarify which decisions remain exclusively under clinician control, and note any contraindications relevant to patient safety and equity. By providing a formalised, bias-aware overview of the AI system, the dossier supports transparency, internal governance, and regulatory oversight, ensuring that pre-deployment decisions explicitly account for risks of inequitable outcomes.

- **A 'combined FRIA and DPIA package'.** Public hospitals deploying high-risk AI systems are increasingly required to assess risks to fundamental rights as well as risks to personal data protection. Integrating these assessments into a single template that satisfies the provisions of the AI Act (Art. 27), GDPR, and DPIA can reduce administrative burden while ensuring that key questions about affected populations, potential harms, mitigation strategies, and complaint mechanisms are systematically addressed, with particular attention to potential inequalities affecting patient populations differentiated by sex and gender and race/ethnicity.

  The 'combined FRIA and DPIA package' operates primarily as an ex ante and periodically updated audit artefact, situated at the pre-deployment stage and within the equity and rights audit layer. These assessments are organised, documented evaluations of risks to fundamental rights and data protection, functioning as governance checkpoints rather than continuous monitoring tools. While they should be updated if risks or system conditions change, their core role is to provide a systematic and reviewable assessment of potential impacts before and during deployment, supporting compliance with both the AI Act and the GDPR, and helping to identify and mitigate gender- or race-related biases.

  To strengthen the human-rights aspect, meaningful consultation with CSOs should be foreseen in impact assessment design to incorporate their rights expertise. For example, the European Network of National Human Rights Institutions (ENNHRI) urges the EC to complement templates with guidance and public consultation to ensure effective FRIAs that consider equity across gender and racial/ethnic dimensions.[41]

- **A 'change-control policy'.** This policy governs updates to AI systems, including model retraining, software upgrades, or modifications to clinical workflows. Organisations should clearly define what constitutes a 'material change' that requires revalidation or approval, ensuring that updates do not introduce unforeseen risks to patient care. In particular, changes must be evaluated for potential impacts on gender and racial equity, including shifts in disaggregated performance or differential outcomes across populations differentiated by demographic characteristics. Hospitals should also ensure that clinicians and relevant staff are informed of modifications that could affect

[41] European Network of National Human Rights Institutions (ENNHRI), *Statement on Ensuring Effective Fundamental Rights Impact Assessments (FRIAs) under the EU AI Act*, Brussels: ENNHRI, 2025.

care outcomes for diverse patient populations, enabling appropriate oversight and mitigation measures. By embedding equity considerations into change management, hospitals maintain both the safety and fairness of AI-supported decision-making in clinical practice.

- **A 'quality management system' (QMS) for deployers who are also providers**. Another important dimension of the model concerns AI developed internally by hospitals or research institutions. Public healthcare organisations increasingly build or adapt their own AI tools, particularly in academic medical centres. In such cases, the institution may effectively act as a manufacturer or provider, which introduces additional regulatory responsibilities. European guidance on in-house medical devices clarifies that healthcare institutions ought to operate quality management systems and maintain technical documentation demonstrating that internally developed tools address specific patient needs that cannot be met through commercial products.[42] The QMS should ensure that AI outputs are evaluated for differential performance across patient populations differentiated by sex and gender and racial/ethnic background, that mitigation strategies are documented, and that human oversight arrangements explicitly preserve patient safety and fairness. Even where regulatory exemptions for in-house devices apply, obligations under fundamental rights and broader AI governance frameworks remain fully relevant, making the QMS a critical tool for embedding equity into the design, validation, and deployment of biomedical AI systems.

Monitoring, auditing and accountability stage

- **A 'population-disaggregated performance and equity report'.** Because variations in algorithmic performance can contribute to unequal health outcomes, hospitals should not only require that suppliers provide relevant information prior to procurement but should also produce internal analyses showing how systems perform across patients with varying demographic characteristics. Those analyses are to focus on stratified performance by sex and gender and relevant racial or ethnic differences where lawful and feasible, as well as justification of any unavoidable performance gaps. Where performance gaps exist, organisations should document mitigation strategies and monitor the effectiveness of those measures over time. This approach reflects emerging expectations within European AI regulation that organisations actively assess and mitigate bias risks in datasets and model outputs. The AI Act (Art. 10) explicitly requires examination and mitigation of bias risks in datasets, and permits limited processing of special-category data for bias detection under strict safeguards.

  The 'population-disaggregated performance and equity report' functions primarily as a continuous monitoring tool within the governance framework. It requires ongoing data collection, regular tracking of system performance across relevant demographic

[42] Medical Device Coordination Group, *MDCG 2023-1 – Guidance on Health Institution Exemption (MDR/IVDR)*, Brussels: European Commission, 2023.

groups, and continuous updating of mitigation measures where differential outcomes are identified. This enables healthcare organisations to detect and address bias risks over time. At the same time, the key findings of this report, such as identified disparities and the effectiveness of mitigation strategies, should be synthesised and included in the 'equity and rights audit', ensuring transparency and public accountability.

- **A two-layer audit regime comprising a technical audit and an equity and rights audit.** A feasible audit regime for hospitals should balance rigour with practicality and be understandable to both staff and the public. In this context, an audit is a formalised, periodic review of an AI system's compliance, performance, and impacts on rights, designed to complement ongoing monitoring rather than replace it.

  Technical audit: this layer constitutes a confidential operational review and covers detailed checks that remain internal, including security testing, robustness assessments, clinical performance validation, and data governance compliance. Particular attention should be paid to performance across sex and gender and racial or ethnic differences, evaluating whether model outputs systematically disadvantage any population. These activities should align with relevant provisions of the AI Act (Arts. 9-15), the MDR and the IVDR where applicable.

  Equity and rights audit: this second layer focuses on transparency, accountability, and bias mitigation. It should document outcome inequalities, the populations included in testing, known limitations, and complaint mechanisms, with explicit analysis of gender and racial impacts. Evidence that human oversight is meaningful and able to correct biased outputs should also be captured. Following the 'bias audit & disclosure' approach used in employment regulation,[43] a summary targeting the general public can communicate key findings, mitigation strategies, and improvements implemented, without revealing proprietary or confidential information. The 'combined FRIA and DPIA package' is also situated within the equity and rights audit layer, reinforcing rigorous evaluation of fundamental rights, data protection, and demographic equity.

- **An 'AI-vigilance log'.** Post-deployment monitoring is also supported through the creation of an AI vigilance log, inspired by pharmacovigilance systems used in medicines regulation.[44] Healthcare professionals should be able to report incidents, near-misses, or unexpected outcomes involving AI-supported decisions, with explicit fields capturing any potential sex and gender or racial/ethnic disparities observed. Reports should be systematically analysed for emerging patterns, including differential impacts on specific demographic groups, enabling organisations to detect safety signals, equity concerns, or bias trends. Where necessary, incidents, in

[43] New York City Department of Consumer and Worker Protection, *Automated Employment Decision Tools (AEDT)*, New York: City of New York, 2025.

[44] European Medicines Agency, *GVP Module VI – Collection and Submission of Suspected Adverse Reactions (Rev 2)*, London: EMA, 2017.

particular those indicating inequitable treatment or risk to protected groups, should be communicated promptly to system providers, relevant hospital governance committees, and regulatory authorities. The log should also feed into periodic equity and rights audits, supporting continuous improvement and mitigation of bias in clinical AI deployment.

- **Network-level collaboration and incident sharing.** The European network of CSOs and public hospitals developed under the AEQUITAS framework functions as a shared governance layer, providing a common vocabulary and reporting standards for incidents and findings related to gender and racial bias in AI systems. This network supports cross-site learning and the dissemination of best practices for detecting, mitigating, and remediating inequities, similar to approaches in high-safety sectors and pharmacovigilance. By linking data on bias incidents, disaggregated performance, and mitigation strategies, the network enables hospitals and CSOs to identify systemic patterns of inequity and coordinate targeted interventions. As the EHDS facilitates both primary use (care delivery) and secondary use (research and policy) under a regulated framework, the network can align its data-sharing and learning practices with EHDS timelines and implementing acts as they mature, supporting more equitable and transparent AI deployment across institutions.
- **Training module for clinicians and CSOs experts.** To mitigate automation bias and inequities in clinical AI, the AEQUITAS regulatory model includes mandatory training designed to equip healthcare professionals with the skills to understand AI system limitations, interpret outputs critically, and override or stop systems when outputs may produce biased or inequitable outcomes. The training emphasizes awareness of gender and racial bias risks, fundamental rights principles, and ethical considerations, including equity, non-discrimination, and patient inclusion.

  This module could be viewed as an enabling artefact, actively supporting the use of the regulatory model by ensuring staff responsible for AI oversight can competently implement monitoring, validation, and accountability procedures with an equity lens. The training is to include case-based exercises on detecting and mitigating sex and gender and racial/ethnic disparities.

Conclusion

AI offers significant opportunities to enhance healthcare outcomes, but without careful governance, it can reproduce or aggravate sex and gender and racial or ethnic inequities, as well as other forms of systemic bias. The regulatory model put forward in this policy document translates fundamental rights principles into concrete safeguards across the AI lifecycle, from procurement and validation to deployment, monitoring, and accountability, with explicit attention to detecting and mitigating gaps in outcomes across varying patient populations. Through implementing an AI governance architecture structured around a clear set of lifecycle gates, policies, practical tools, and artefacts, hospitals and CSOs can systematically assess, mitigate, and document risks, including inequitable impacts on marginalised patient populations. By integrating cross-sector lessons, EU legal requirements,

and international governance frameworks, the regulatory model supports healthcare professionals in administering AI systems responsibly and ensuring equitable, safe, and transparent healthcare for all patients, regardless of sex and gender, racial/ethnic origin, or other characteristics.