CORE-MD

Coordinating Research and Evidence for Medical Devices

Expert advice on criteria for the regulatory evaluation of ML and AI

D2.4

Deliverable factsheet

Source Activity:	Work package 2, Task 2.3
Title:	Expert advice on criteria for the regulatory evaluation of ML and AI
Lead Beneficiary:	KU Leuven
Nature:	Report
Dissemination level:	Public
Editor:	Frank E. Rademakers (KU Leuven)
Authors:	Frank E Rademakers (KU Leuven), Elisabetta Biasin (KU Leuven), Bart Bijnens (KU Leuven), Nico Bruining (Erasmus MC)*, Enrico G. Caiani (POLIMI), Koen Cobbaert (Philips)*, Rhodri H. Davies (University College London)*, Job N. Doornberg (University Medical Center Groningen)*, Stephen Gilbert (Technische Universität Dresden), Leo Hovestadt (Elektra), Erik Kamenjasevic (KU Leuven), Zuzanna Kwade (Dedalus)*, Gearoid McGauran (HPRA), Gearoid O'Connor (HPRA), Baptiste Vasey (UOXF) and Alan G Fraser (ESC) **External experts involved in CORE-MD activities*
Status:	Final
Date:	12/04/2023
Contractual Delivery Date:	Month 24

Version Log

Issue Date	Version	Involved	Comments
15.11.2022	V1	Frank Rademakers (KU Leuven)	First version
21.12.2022	V2	Frank Rademakers (KU Leuven) and task 2.3 members	Inputs from task members
08.03.2023	V3	Frank Rademakers (KU Leuven) and task 2.3 members	Inputs from task members
27.03.2023	V4	Frank Rademakers (KU Leuven)	Consolidation
28.03.2023	V5	Alan Fraser (ESC)	Coordinator's quality check
12.04.2023	V6	Jean-Baptiste Rouffet (EFORT) | Valentina Tageo (ESC)	Final editing for submission

Acronyms and abbreviations

AI	Artificial Intelligence
ACP	Algorithm Change Protocol
BRA	Benefit-risk Assessment
CE	Communauté Européenne
CEN	Centre Européen de Normalisation – European Standardisation Center
CPS	Clinical Performance Score
EU	European Union
FDA	Food and Drug Administration (USA)
DL	Deep Learning
GDPR	General Data Protection Regulation
GMLP	Good Machine Learning Practice
HCP	Healthcare professional
HTA	Health Technology Assessment
IMDFR	International Medical Device Regulator Forum
ICHOM	International Consortium for Health Outcome measurement
IEC	International Electrotechnical Commission
ISO	International Standardisation Organisation
IVD	In Vitro Diagnostic
MDCG	Medical Device Coordination Group (EU)
MDR	Medical Device Regulation
MDSW	Medical Device Software
MHPRA	Medicines and Healthcare Products Regulatory Agency
ML	Machine Learning
NB	Notified Bodies
NICE	National Institute for Health and Care Excellence
NIST	National Institute of Standards and Technology (USA)
PMS	Post-market Surveillance

SHAP	Shapley Additive Explanations
SAMD	Sofware as Medical Device
SPS	Software Pre-Specification
QMS	Quality Management System
TEVV	Test, Evaluation, Verification and Validation
TPS	Technical Performance Score
VCAS	Valid clinical association score

Table of Contents

Index of figures

Index of tables

Executive Summary

In the healthcare setting, providers use evidence-based guidelines in care programs and pathways that aim to standardize the diagnostic and therapeutic approach to a specific medical problem. Healthcare professionals may be informed by guidelines provided by scientific societies that abide by standards and rules, including conflicts of interest. The overall principle in health care, "*primum non nocere*", "first, do no harm," also applies to AI tools. The balance between positive outcomes and possible safety risks and side-effects, both at the individual and the societal level, needs to be considered to define the evidence required to demonstrate the existence of benefit on the outcome and workflow against the background of a potentially significant negative impact on human rights. Individual and societal human rights might conflict to a certain degree, and ethical considerations must prevail in these circumstances.

Artificial intelligence covers a wide variety of digital tools, ranging from (relatively) simple algorithms to deep learning methods. The range of autonomy of AI systems and the degree of possible (human) supervision vary significantly. Several AI tools have also become available as apps to be used by citizens and patients without any involvement of healthcare professionals, in which case oversight depends entirely on the end user's appreciation.

The potential of AI tools for the implementation of personalized medicine depends mainly on adapting the tool to the specific use setting and to its end-user's personal characteristics and history. This adaptation requires a self-learning or adaptive approach with the continuous or intermittent implementation of changes, adding to the complexity of the post-release phase and making Algorithm Change Protocols challenging to implement.

Just like for any other medical device, the availability of an AI tool will depend on it having satisfied market authorization frameworks proving its safety and performance and also market access frameworks weighing its cost against its value. Market authorization and access frameworks require a certain degree of transparency so that independent parties (Notified Bodies (NB)) can evaluate the AI tool and the processes used to design and develop it. Good documentation and record-keeping practices involving auditability and traceability facilitate this transparency. Additionally, healthcare professional look for guidelines or clinical evidence to support the use of the AI tool in specific circumstances.

As the potential pervasiveness of AI tools can threaten their users' well-being, many official entities have called for strict(er) regulation. Many initiatives have been taken to address this point or are underway, in Europe and elsewhere. The problem for end-users persists as most of these initiatives focus on the underlying principles, but practical guidance remains scarce.

This document aims to provide a roadmap for practically implementing the many laws, regulations, consensus recommendations, and guidance documents for the clinical evaluation of AI for healthcare purposes, while identifying potential improvements for future legislative revisions. As AI covers many tools, this paper adopts a risk-based approach to create a requirement matrix for the clinical evaluation of these devices in the pre-release and post-release phases. Emphasis is on the tool's clinical usefulness,

which must consider the real-life implementation and integration into care pathways. For an AI medical device to be successfully implemented, it is essential that it can be demonstrated that the tool's outputs are reliable and trustworthy and that its claimed clinical benefits are substantiated, so that health professionals, end-users, and society as a whole can accept its use. A post-release evaluation should, therefore, not only deal with relevant clinical outcomes for the end-users but also with the broader impact of the tool on individuals, society, the environment and the planet, while considering the tension between safeguarding individual human rights and improving general societal benefit. Further education on AI tools to enhance the skills of healthcare professionals and end-users is essential in this context, not only for the correct use of the AI tool but also, more generally, to increase acceptance and trust by the general public.

1. Introduction

Artificial Intelligence (AI) covers a wide variety of digital tools, ranging from (relatively) simple algorithms to deep learning methods [1][2]. AI is increasingly proposed for healthcare applications that support healthcare providers (HCP) in formulating a diagnosis [3], predicting the course of disease [4][5], selecting treatment, managing patients, monitoring outcomes, and supporting shared decisions with patients and healthy citizens [6][7][8][9][10][11] . These applications are referred to as "decision support systems", and are generally considered among the more complex and critical AI tools which require clinical evaluation. They are therefore the focus of this paper [12].

It delivers D2.4 associated to the Task 2.3 “Developing guidance for the evaluation of AI and standalone software in medical devices” in WP2 of the CORE-MD project and formulates possible next steps to the clinical evaluation of high-risk Medical Device Software (MDSW) using AI. As it focuses on AI use in MDSW it overlaps and interacts with many devices using or relating with such software and complements the work in WP1 and Task 2.1 in WP 2.

The range of autonomy of AI systems and the degree of possible (human) supervision vary significantly. Most observers [13][14][15] stress the need for human oversight and final decision-making by a healthcare professional[15][16][17]. They underline the need to integrate AI tools into the current workflow and create a ‘Human–AI team’. The very nature of some of the AI tools, however, makes interpretation of AI results and oversight difficult, because they perform as black boxes [18], and the reasoning behind the ultimate result/suggestion/decision of the AI tool remains obscure, if not lacking, even with the use of explainability methods [19][20][21][22]. The effectiveness of oversight is reduced for less experienced users, who, paradoxically, might benefit the most from such tools [23]. For some applications, providing real-time human oversight might even reduce safety and decrease the performance of the AI tool, e.g., when it has been shown experimentally that the AI tool exceeds the capabilities of the human and the human-AI team, both in terms of speed (fast reaction times), performance (accuracy and precision), or being less prone to mistakes due to intrinsic human factors. Several AI tools have also become available as apps to be used by citizens and patients without any involvement of healthcare professionals, in which case oversight depends entirely on the end user's appreciation[24].

Finally, the potential of AI tools for the implementation of personalized medicine depends mainly on adapting the tool to the specific use setting and to its end-user's personal characteristics and history. This adaptation requires a self-learning or adaptive approach with the continuous or intermittent implementation of changes, adding to the complexity of the post-release phase and making Algorithm Change Protocols challenging to implement[25]. Besides such structured self-learning and personalization, with additional data collection being planned and used for algorithm adaptation, a more or less detectable drift or enlargement in the target population, the intended use, or the human-AI interaction could all change the risk and performance metrics of the AI tool, necessitating continuous evaluation after its release. These needs support a more agile approach in the development, testing, and

validation of AI tools (The Last Mile: Where Artificial Intelligence Meets Reality. Enrico Coiera) and a system approach rather than a pure device focus[26][27][28].

As the potential pervasiveness of AI tools can threaten their users' well-being[29][30][31][32][33], many official entities have called for strict(er) regulation[34][35][36][37][38][39]. Many initiatives have been taken to address this point (MDR, Data Governance Act) or are underway, in Europe (AI Act [40][41][42], AI Liability Directive Proposals) and elsewhere (IMDRF[43] (globally), USA [44][45][46][47][48][49][50], China [51], Canada [52], and several other countries nationally [53][54][55][56]**¡Error! No se encuentra el origen de la referencia.**[58][59][60]). The problem for end-users persists as most of these initiatives focus on the underlying principles, but practical guidance remains scarce.

1.1 Deliverable structure

This paper aims to provide a roadmap for practically implementing the many laws, regulations, consensus recommendations, and guidance documents for the clinical evaluation of AI for healthcare purposes [61][62], while identifying potential improvements for future legislative revisions. Such a roadmap could limit the interpretation variability of manufacturers, Notified Bodies, and reviewers. As AI covers many tools, this paper adopts a risk-based approach to create a requirement matrix for the clinical evaluation of these devices in the pre-release and post-release phases. Emphasis is on the tool's clinical usefulness, which must consider the real-life implementation and integration into care pathways. For an AI medical device to be successfully implemented, it is essential that it can be demonstrated that the tool's outputs are reliable and trustworthy and that its claimed clinical benefits are substantiated, so that health professionals, end-users, and society as a whole can accept its use. A post-release evaluation should, therefore, not only deal with relevant clinical outcomes for the end-users but also with the broader impact (whenever relevant) of the tool on individuals, society, the environment and the planet, while considering the tension between safeguarding individual human rights and improving general societal benefit. Further education on AI tools to enhance the skills of healthcare professionals and end-users is essential in this context, not only for the correct use of the AI tool but also, more generally, to increase acceptance and trust by the general public [62].

After formulating the background to the place of AI MDSW in health care, the document focuses on a risk-based approach and on the specific requirements this entails in the pre- and post-release phases of the life-cycle of a MDSW. Subsequently the different steps of this approach are formulated.

2. Background

Healthcare providers use evidence-based [63] guidelines in care programs and pathways that aim to standardize the diagnostic and therapeutic approach to a specific medical problem. Such standardized approaches can help health practitioners to improve patient outcomes and to improve the health potential of the individual and, ultimately, the population [64]. Decisions to deviate from these approaches are warranted following consideration of the patient's unique background, needs and expectations. The healthcare professional makes many such decisions and has to be able to justify these to people affected by the decision.

The professional formally documents the decision when its impact is significant and~~,~~ when legal obligations require this, or when the professional strives to involve and empower the patient depending on the health-care context [65] (according to Lilrank operational modes: prevention, acute care, one-off/single intervention, focused factory, cure or care). Other circumstances will also influence whether or not to share the context of a decision. For example, when ordering an imaging test, the decision to use a specific imaging machine's brand is seldom discussed with the patient but driven mainly by availability. On the other hand, when choosing between surgery and medical therapy (if those options exist in the guidelines), that decision is extensively discussed by explaining in clearly understandable language the different possible approaches, their pros and cons, the alternatives, and the consequences of taking no action at all.

The healthcare professional may be advised for this discussion by guidelines provided by scientific societies that abide by standards and rules, including conflicts of interest. They use evidence from clinical trials, case reports, and other information, leading to a specific level of confidence for a recommendation.

Similar processes exist for the use of drugs and devices but here the manufacturer also plays a very important role.

Just like for any other medical device, the availability of an AI tool will depend on it having satisfied market authorization frameworks proving its safety and performance (but often in circumstances that are not relevant to clinical conditions) (e.g., CE-marking, 510(k), de Novo, premarket approval), and also market access frameworks weighing its cost against its value (e.g., health technology assessments, NICE, DiGA). Market authorization and access frameworks require a certain degree of transparency so that independent parties (Notified Bodies (NB)) can evaluate the AI tool and the processes used to design and develop it. Good documentation and record-keeping practices involving auditability and traceability facilitate this transparency. Additionally, the HCP looks for guidelines or clinical evidence to support the use of the AI tool in specific circumstances.

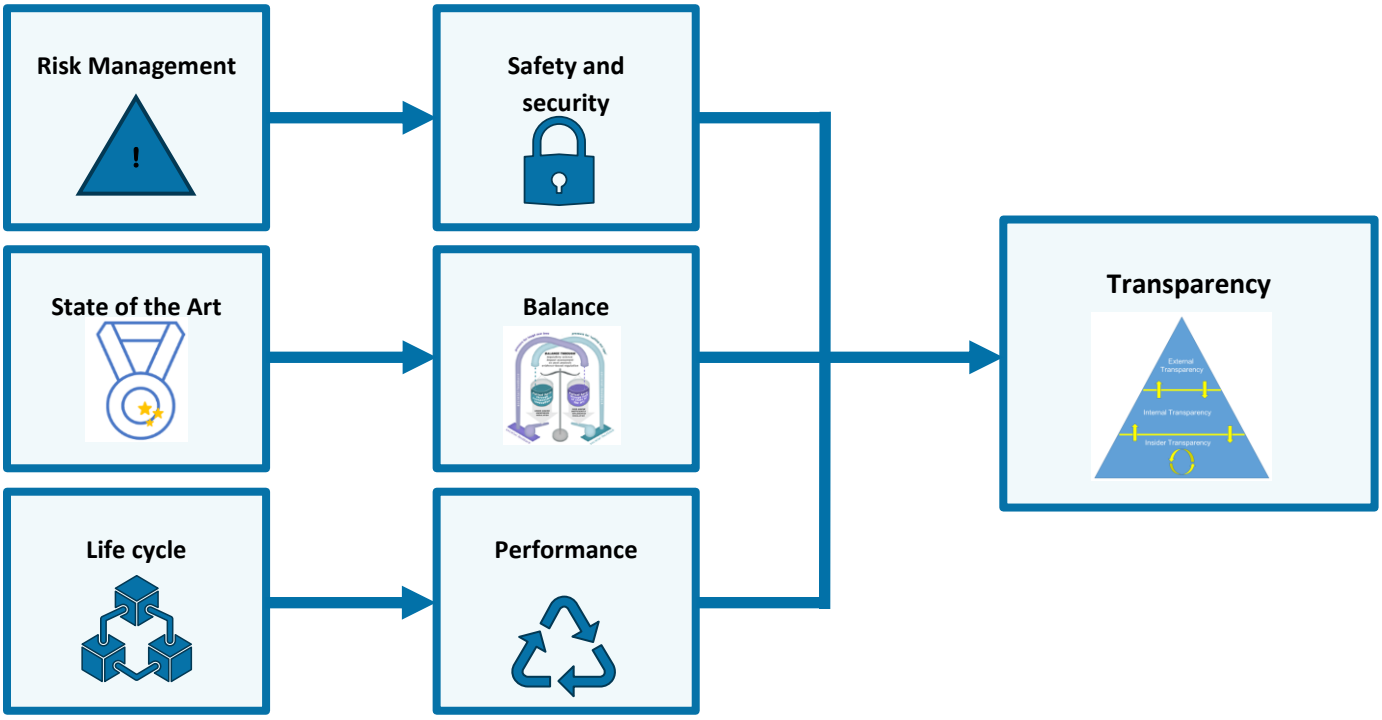

Figure 1. Regulatory Requirements for AI tools

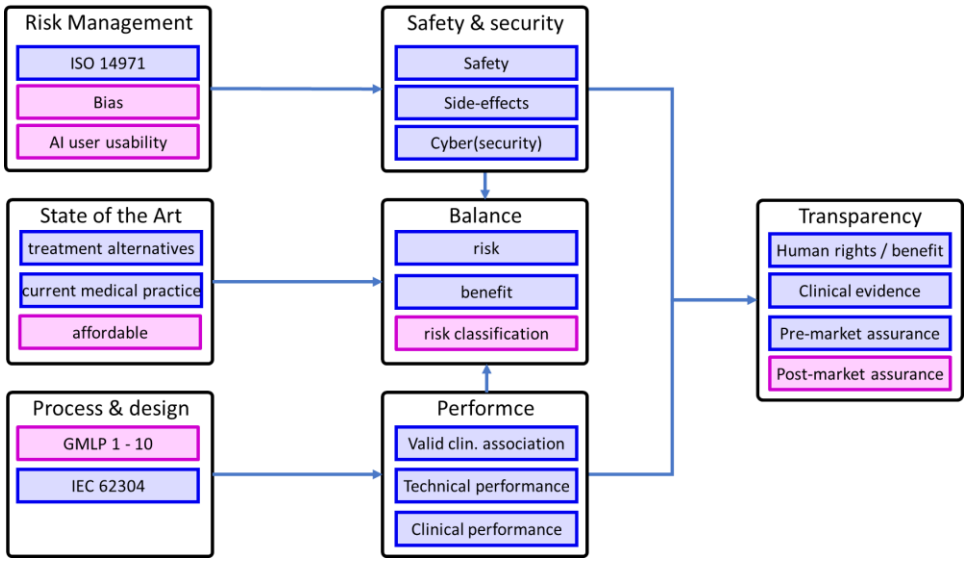

Figure 2. An overview of the high-level regulatory requirements for AI (with areas for possible improvements in standards when compared to current regulations shown in purple)

Transparency and the access to information are core concepts in many recent EU legal acts. The GDPR, for example, protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data [66][67][68]. One important principle of data processing is transparency [69] . Defining the ultimate goal of this principle, i.e. "why" transparency is needed, is imperative to understand the importance of the transparency and its application in healthcare. Part of the transparency aims to balance individual human rights with rights of other individuals and society at large [70], such as for the secondary use of data in healthcare if there is a legitimate or public interest. Also for AI, one should consider the ultimate goal for the AI tool's application in healthcare and the proportionality between risk and benefit of the tool for achieving that specific goal during the entire life cycle of the tool [71].

To achieve that, the HCP can rely on information provided by the AI tool's manufacturer to assess the tool's utility, explain its function, risks and impact, and decide with the patient on its use. If the patient wishes to know, the HCP should be able to explain (explainability) why a certain device/software is used in that specific circumstance and with what advantages and disadvantages, so that the patient can interpret/comprehend (interpretability) the possibilities and can agree (informed consent) or make a choice together with the HCP (shared decision making). This can be explicit or implicit (post-hoc, liability). This dialogue has been named **external transparency** in [72]. To be able to do so the HCP needs professional guidelines and recommendations and information from the manufacturer/provider of the device, including evidence from tests and clinical trials [73]. It is an obligation of the manufacturer/provider to provide this information (explain the device) in a format and at a level that is interpretable by the HCP. This can be called **internal transparency**. Finally for the manufacturer to be able

to provide that explanation to the HCP, **insider transparency** is needed within the company, and is mostly based on standards and procedures by which the manufacturer operates (IEC 62304 (generic) and GMLP(AI specific) [74]).

AI software is no different from other products (software, devices, the use of medication or other therapies) in its obligation to provide transparency at these different levels but its capacity for explainability can vary significantly depending on the specific use of machine learning or deep learning and on the supervised or unsupervised nature of the learning process [75].

In **summary, transparency comprises the following aspects:**

- **Why**
- **What**
- **How**
- **When**

For **transparency between healthcare professionals and patients or citizens (external transparency)**, the constituents are:

- **Why**: enable informed consent/shared decision making
- **What**: inform about purpose and nature of AI tool's intervention, consequences and risks, performance/chance of success
- **How**: adapt to the needs, expectations and literacy of the patient
- **When**: before an intervention (depending on interaction patient – HCP)

An AI tool built using specific software techniques like deep learning differs from other healthcare tools in its capacity to provide transparency, due to the difficulty of interpreting its underlying function. **Interpretability** is thus an important constituent of transparency (besides accountability, auditability, traceability, documentation, reporting, …) and can be achieved by **explainability** but also by testing results and their contextualisation, by verification (by HCP of the result of the AI tool) or by experimental evidence.

In the case of AI decision-support systems, external transparency consists of notifying the patient that an AI device is used and to what degree it contributes to the overall decision, and of describing the alternatives to that use with their advantages and disadvantages, the level of confidence (if the AI tool can be explained), and whether it has been verified or has been clinically tested, if it has undergone certification, and what the level of accuracy is The patient should also be aware what specific personal data has been used in the AI tool

Legislation for all medical devices, even when their explainability is low (opaque systems), requires that testing and clinical evidence must be appropriate to elucidate the benefit/risk ratio for indicating its use in specific circumstances. In cases where the evidence is inappropriate and/or the benefit/risk ratio is low,

its use must not be indicated. For minority populations and uncommon indications, the benefit/risk ratio is an important co-determinant of the required level of evidence.

While this external transparency between patient/affected person and HCP is not an explicit requirement under the MDR (but falls under the medical legislation in many member countries), from a clinical perspective it is the logical reason to achieve internal transparency between manufacturer and HCP, which is a core part of the MDR. It connects the more scientific, clinical approach to evidence-based decision making, with the specific requirements under the MDR, in a way that is understandable to all parties involved. It does imply that evaluation of the internal transparency should involve the HCP's who will be using the device in clinical practice, and it requires extension of the evaluation of impact of the AI tool to all affected persons or groups (art 3 AI Act).

For **transparency between manufacturers and healthcare professionals (internal transparency)**, the constituents are:

- **Why**: enable HCP to explain the use of the AI tool to patients/citizens
- **What**: inform about performance, accuracy of AI tool; data used to train and test; results of verification and validation; existing experience and real-life know-how
- **How**: adapt to the needs, expectations and literacy of the HCP
- **When**: before use of the tool by HCP

For insider transparency IEC 62304 describes the software life cycle process, however IEC 62304 does not contain specific requirements for Artificial Intelligence. The Good Machine Learning Practices (GMLP) set up by the FDA / Health Canada and MHRA describes those requirements for AI. The following figure shows the integration of IEC 62304 and the GMLP principles. The GMLP principles are given in red. The GMLP principles adds the missing AI requirements. (further relevant ISO standards are ISO 14971:2019 and ISO 13485:2016).

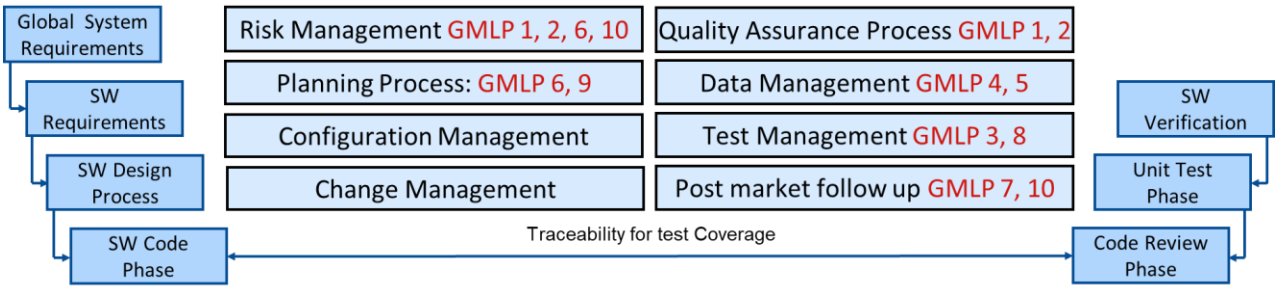

Figure 3. IEC 62304 describes the software life cycle process

The GLMP's are:

1. Multi-Disciplinary Expertise Leveraged Throughout the Total Product Life Cycle
2. Good Software Engineering and Security Practices Are Implemented
3. Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population
4. Training Data Sets Are Independent of Test Sets
5. Selected Reference Datasets Are Based Upon Best Available Methods
6. Model Design Is Tailored to the Available Data and Reflects the Intended use of the Device
7. Focus Is Placed on the Performance of the Human-AI Team
8. Testing Demonstrates Device Performance during Clinically Relevant Conditions
9. Users Are Provided Clear, Essential Information
10. Deployed Models Are Monitored for Performance and Re-training Risks Are Managed

The three general principles of the MDCG 2020-1 recommendation also fit with these transparency levels: valid clinical association corresponds mostly to the external transparency; clinical performance relates to the internal transparency; technical performance coincides with the insider transparency level [76][77].

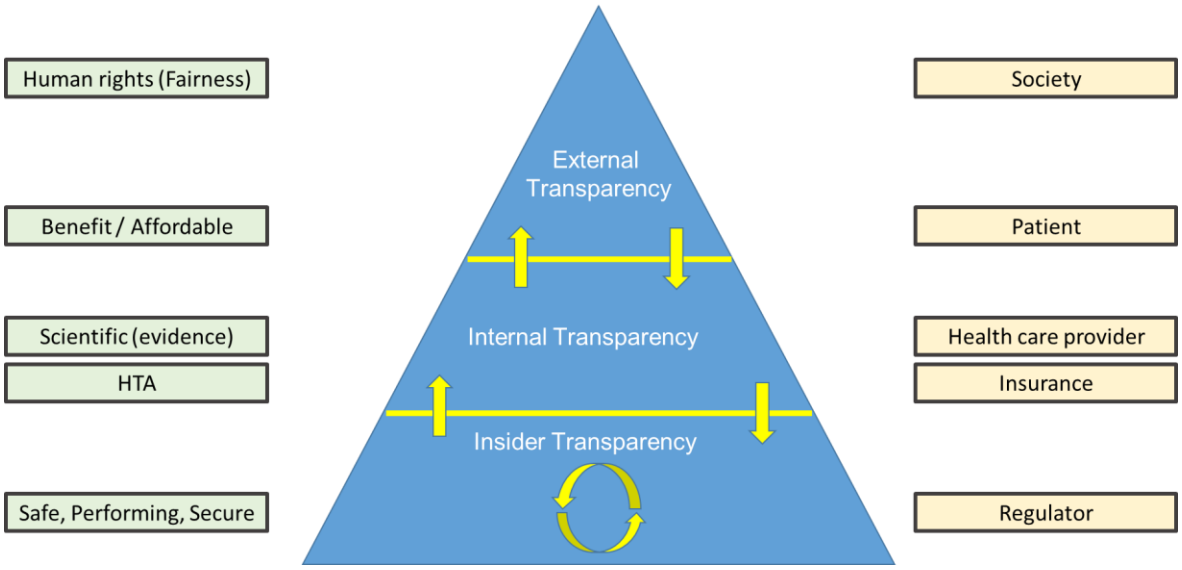

Figure 4. Transparency relation to MDCG 2020-1 recommendation

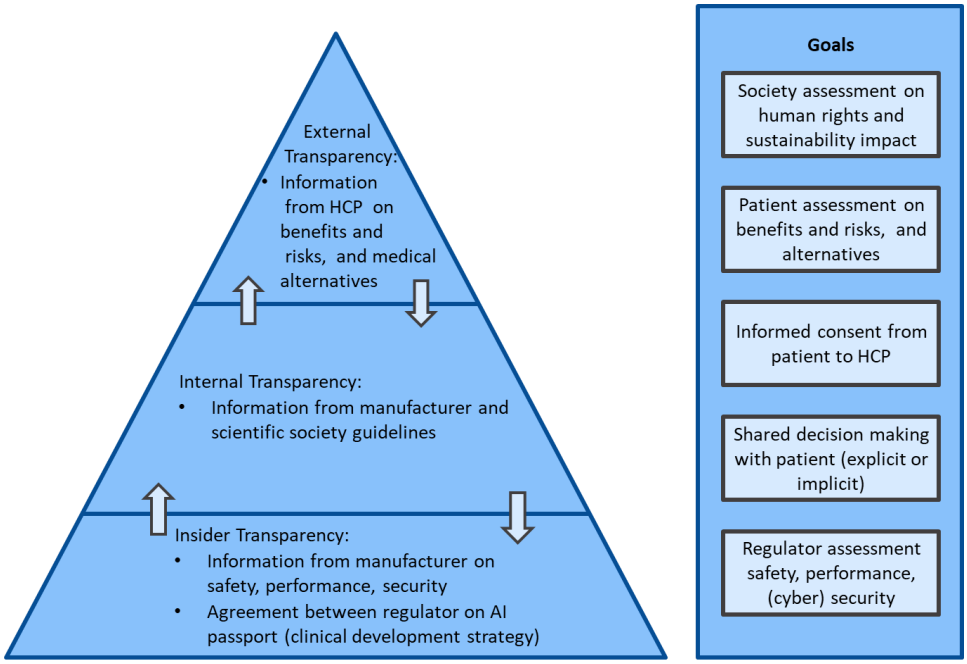

Figure 5. Transparency levels and associated goals

3. Risk-based Approach

The overall principle in health care, "primum non nocere", "first, do no harm," also applies to AI tools. The balance between positive outcomes and possible safety risks and side-effects, both at the individual and the societal level, needs to be considered to define the evidence required to demonstrate the existence of benefit on the outcome and workflow against the background of a potentially significant negative impact on human rights. Individual and societal human rights might conflict to a certain degree, and ethical considerations must prevail in these circumstances [78]. Medical device legislation stresses this risk-based approach, requiring evidence for a positive benefit-risk balance, including evidence of safety and performance in consideration of the state-of-the-art. That level of evidence must be appropriate in view of the characteristics of the AI tool and its clinical use. Manufacturers are required to justify why the level of clinical evidence provided is sufficient to meet conformity standards while considering three aspects crucial for its safe and effective use: valid clinical association, technical performance, and clinical performance [76].

Risk refers to the composite measure of an event's probability of occurring and the magnitude or degree of the consequence of the corresponding event. So, for AI tools, risk is a function of 1) the negative impact or magnitude of harm that would arise if the circumstance or event occurs and 2) the likelihood of occurrence. Risk management is relevant at every lifecycle stage of an AI tool, to lead to trustworthy AI systems that deliver their potential benefits to people (individuals, communities, society), organisations and systems. Specific challenges for AI risk management include the difficulty: 1) to define and measure AI risks and failures; 2) to deal with risk tolerance which is subject to legal and regulatory requirements but also depends on shifting organizational and societal acceptances and preferences; 3) to prioritise risk by considering the absolute risk but also the risk culture of the use environment and the residual risk; 4) to allow for variable risk in view of the organizational integration at the different steps of the AI lifecycle and the management of other risks such as cybersecurity and privacy.

While the AI actors vary across the AI lifecycle stages, at every stage the goal is to optimize the ultimate risk-benefit balance for the end-user(s) by including the relevant stakeholders and to provide the optimal TEVV (test, evaluation, verification and validation) processes. These different stages can be grouped as Data and Input; AI Model; Task and Output; Application context (NIST) [79][80][81][82][83]. Trade-offs will usually be necessary between the many typical characteristics of trustworthy AI: ideally it should be valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed (NIST) [84].

A specific characteristic of many AI tools is the need for high-quality training and testing data sets [85][86] and, in the case of supervised learning, relevant labels, i.e., characteristics determined by qualified human annotators for use as a reference standard (ground truth). The collection and use of such data sets, when containing personal information, are subject to the requirements of the Data Protection Regulations [87][88]. The specific use of such data sets for developing AI tools and the need for curation necessitates extra attention concerning their validity, intrinsic bias (i.e., by race, gender, age group, etc.),

representativeness in terms of the group of patients under focus, the geographic distribution of data sources, quality of the labels (i.e., level of experience of the medical operators, their qualifications, their geographic distribution). Despite increased awareness and additional efforts, the chances are still high that the device will ultimately be used in a population with slightly other characteristics than the one used in the training and testing sets, resulting in performance drift. As such, a stronger emphasis is needed on post-market surveillance, as only long-term monitoring in routine clinical use can provide the necessary evidence of accurate clinical performance.

Providers of AI tools with a favourable clinical benefit-risk ratio presenting a low risk to individuals and society could release such tools with evidence powered for less stringent (statistical) significance or generalizability from the premarket phase and a more significant emphasis of data and evidence gathering in the post-market phase. For instance, a manufacturer can release an AI tool with evidence powered ~~for~~ with statistical significance for the general target patient population, but not for all subgroups. The manufacturer will then need to caution for possible suboptimal performance in these subgroups but may not wish to exclude these populations to not endanger availability of adequate diagnostic or treatment solutions for minority populations, taking risk-benefit into consideration.

In contrast, tools with an unfavourable benefit-risk balance and a high risk to individuals or society should undergo an extensive pre-release clinical evaluation phase, including medical investigations/trials.

In such a scenario, appreciation of possible benefits and risks is essential to decide how pre-release and post-market requirements should be fulfilled (see next chapter).

Several factors, including already existing evidence, should be taken into account for the **risk-benefit assessment**:

1. **Medical Purpose**
    a. prevention,
    b. screening/prediction,
    c. diagnosis,
    d. treatment,
    e. monitoring,
    f. workflow recommendation,
    g. …
2. **The intended medical conditions**
    a. symptoms, disorders (genetic, …), diseases,
    b. their stage/level of severity,
    c. seriousness of manageable outcomes;
3. **The intended population**
    a. healthy individuals/patients,
    b. neonatal/infants/children/adults/elderly,
    c. hospitalized/ambulatory,

d. acute/chronic,
e. …

4. **The AI tool's operation**
   a. inputs: data
      i. known/unknown, type (synthetic?), origin, acquisition environment
      ii. risk for bias
      iii. curated by? expertise
      iv. privacy and security
      v. ...
   b. Algorithm
      i. supervised/unsupervised
      ii. ground truth, guideline, reference standard possible/impossible and known/accurate
      iii. algorithmic, ML, DL
      iv. degree of learning/ change management autonomy
         1. self-learning, batch training
         2. lay user/non-clinical user driven
         3. clinical user driven
         4. manufacturer driven
         5. none/static
   c. Outputs
      i. Flow
         1. automated
         2. semi-automated: approval requested depending on risk
         3. manual
      ii. possible impact/harm of erroneous output: physical and mental
         1. critical/life-threatening
         2. serious
         3. minor
         4. negligible
         5.
      iii. testing, evaluation, validation, verification (TEVV): performance
         1. Parameters used
         2. Required level of parameters
         3. Use of reference labs/datasets
         4.
   d. Presentation of results
      i. not explained/explainable
      ii. partially explained (saliency maps, data visualization, SHAP, …)
      iii. interpretable by

1. Explanation
2. Verification

e. integration in workflow:
   i. Type
      1. autonomous (no possibility for human/HCP? in the loop)
      2. supervisable but normally operates without human approval: drive:
      3. Suggest
      4. Inform
   ii. Timing
      1. closed loop
      2. Sequentially
      3. opportunity for correction
   iii. Usability/user experience
      1. potentially adding unnecessary burden of registration on clinicians, distracting them from other clinical tasks
   iv. Cybersecurity

f. outcomes impacted by AI tool
   i. measurable, quantifiable taking into account the required effort for obtaining these parameters
   ii. intended level of improvement: benefit
      1. cf Porter Hierarchy
      2. ICHOM Prom's and Prem's

5. **the intended user in their respective environment (non-clinical, general healthcare, specialist healthcare) and scope (universal, national, regional, site, patient-specific):**
   a. lay person (healthy individual, patient relative),
   b. patient,
   c. HCP non-medical,
   d. general practitioner,
   e. medical specialist;
6. **The potential shift/drift in**
   a. intended use
      i. Contraindication
   b. intended user
   c. intended environment

For a given AI tool, some or all of these factors will be relevant but usually with variable impact on the overall risk-benefit balance. Therefore the manufacturer needs to position the AI tool with respect to all these factors so that the reviewer/NB (and ultimately the end-user to decide on use of the tool in the clinical environment) can appreciate the risk-benefit balance and decide on the required

information/evidence required pre-release to allow release of the tool and to specify the post-release requirements, i.e. providing an AI tool's Benefit-Risk Assessment (BRA).

The CORE-MD project does not consider whether the AI tool manufacturer proposes to have the tool's output validated by a human, either a healthcare professional or an end-user. The capability of the "human-in-the-loop" to validate the result of the AI tool may depend too much on the end user's expertise and experience. Therefore, the suggestion for output verification is not enough to designate an AI tool as low Risk. Risk appreciation must also consider real-life use and drift[90] in data, model, concept, intended use and users. So, CORE-MD proposes to consider all these factors as part of the risk assessment, which will determine the requirement matrix in both the pre-release and the post-market phase. The post-market phase mandates recurrent risk assessment at a frequency depending on the initial Risk and possible drift.

We propose using a simple point scoring system to determine the overall Risk of an AI tool and direct the level/depth of the pre- and post-release requirements. From a clinical perspective and starting from the MDCG 2020-1, three parts of evidence are required:

1. valid clinical association,
2. technical performance,
3. clinical performance.

For each we propose scores from 1 to 3, relevant to the characteristics of the AI tool and its application, where lower values are associated with lower risk/higher benefit. Although this approach could seem like an oversimplification, it does not deflect from the need for appraisal of the entire AI BRA by the reviewer/NB but it is intended to help manufacturers, NBs and clinicians to prioritize efforts and avoid limiting access to clinical use for potentially helpful AI tools.

Valid clinical association / Scientific validity VCAS

Valid clinical association / Scientific validity is defined as "the extent to which the MDSW's output (e.g., concept, conclusion, calculations), based on the inputs and algorithms selected, is associated with the targeted physiological state or clinical condition. This association should be well founded or clinically accepted" [76].

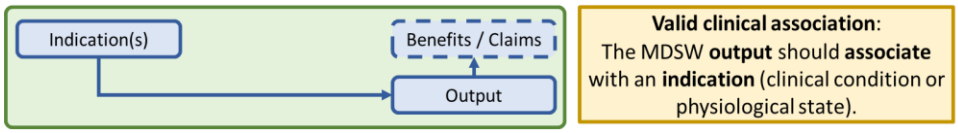

Figure 6. Schematic view of valid clinical association

Such an association is characterized by the type of AI model (algorithmic, Machine Learning (ML) or Deep learning (DL)), the availability or not of a ground truth (supervised or non-supervised) to train and test the algorithm and the resulting transparency, explainability and possibility for human oversight. Such oversight is only relevant when interpreted as the possibility for a human end-user, either citizen/patient or healthcare provider to verify the output of the AI tool against state-of-the art knowledge for that end-user with respect to the intended use of the tool. As an example, a diagnostic imaging segmentation tool

could be using a black-box deep-learning algorithm but the output, i.e. the contour, can be verified by the clinician and validated if it conforms with the state-of-the art knowledge and expertise of the clinician.

This corresponds to point 4b of the AI BRA and is the main determinant of the VCAS.

Table 1. VCAS scoring

Transparency and Oversight	Valid clinical association score (VCAS)
Easy	1
Difficult	2
Impossible	3

Technical Performance Score TPS

Technical performance is defined as the "Capability of a MDSW to accurately and reliably generate the intended technical/analytical output from the input data"[90][91][76].

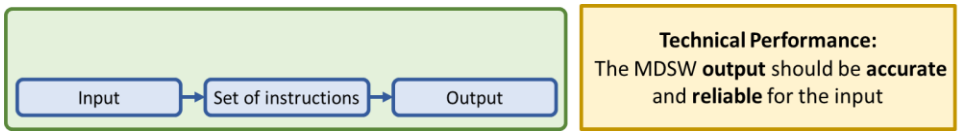

Figure 7. Schematic view technical performance score

Verification of technical performance is thus the demonstration of the ability of the AI tool to accurately, reliably and precisely generate the intended output, from the input data. in real-world usage; in the intended computing and use environments.

Technical performance can be provided by figures of merit that are usually utilized for AI tools (accuracy, specificity, sensitivity, AUROC, F1 score, …). Independently of the resulting numbers, we propose an association with the different level of testing of the AI tool:

Table 2. TPS scoring

Extension of validation/testing	Level of validation/testing	Technical performance score (TPS)
Internal validation	weak	3
Narrow external validation	moderate	2
Broad External validation	strong	1

In the INTERNAL VALIDATION, the AI tool performance has been evaluated just based on testing data acquired with the same settings (same institution, using the same equipment, interpreted by the same observer as the training group, same group of patients,…) than the training data.

In the NARROW EXTERNAL VALIDATION, the training and testing data are partially differentiated for some of the previous factors, while in the BROAD EXTERNAL VALIDATION a high number and variety in the testing data (i.e., using different equipment, from different centres, at different times, interpreted by different observers, in different patient groups, …) compared to the training data, is introduced.

This score also reflects the risk of introducing a bias in the performance of the AI model results, relevant to the data utilized for training and to the more or less extensive testing on similar data, or in data containing more variation.

This corresponds to points 4a,4c-f of the AI BRA.

Clinical Performance Score CPS

Clinical performance is the "ability of a device, resulting from any direct or indirect medical effects which stem from its technical or functional characteristics, including diagnostic characteristics, to achieve its intended purpose as claimed by the manufacturer, thereby leading to a CLINICAL BENEFIT for patients, when used as intended by the manufacturer". ([76] - only for MD (not IVD))

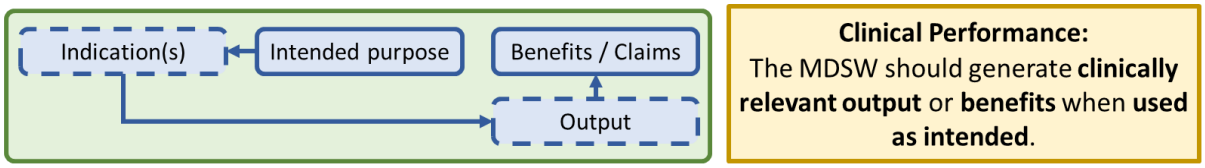

Figure 8. Schematic view clinical performance score

We propose a score associated to the intended purpose (i.e., the intended use claimed by the manufacturer, with respect to the target population and the clinical indication) of the AI tool, also considering the classification system proposed by the IMDRF, which differentiates the level of significance of information (i.e., inform or drive clinical management, diagnose/treat) and takes into account the healthcare situation/condition of label use (i.e., non-serious, serious, critical). Such criticality is, however, very context-dependent. The same disease or condition may be acute or chronic, with various levels of severity, and impacted by co morbidities. An AI tool may drift from the intended purpose and use with respect to the criticality when in real world use and therefore it is best to consider the most critical possible use for determining the risk score at onset. As such a diagnostic tool could be critical when the result determines the choice of treatment of a life-threatening disease, while the same diagnostic tool could be used for a non-serious aspect of a chronic, non-life-threatening illness or with extensive human oversight.

Ultimately, clinical performance can only be determined during real-life use.

Table 3. Clinical Performance scoring

Criterion	Associated Levels	Partial score	Clinical Performance Score (CPS)
Type of disease, condition, disability, healthcare situation: risk for patient			
	Non serious	1	
	serious	2	
	critical	3	
Significance of information: use in clinical flow	Inform	1	
	Drive	2	
	Diagnose or treat	3	
			Sum of the two

The CPS refers to the criticality of the healthcare situation for which the AI tool is intended, as well as the expected impact of the output information in the context of the clinical workflow. This could be entirely determined by the intrinsic nature of the AI tool itself but more often would be defined by the use made of the tool during clinical implementation. Such use is specified by the manufacturer but drift can occur with or without the knowledge of the manufacturer ("off-label" use) and this again stresses the need for post-market surveillance of real-life clinical use and its associated risk, which can warrant additional testing and validation.

This corresponds to points 1-3, 5-6 of the AI BRA.

To avoid inappropriately release of an AI tool with a low overall score but a high score on one of the subsets, we propose a flow chart which subsequently considers these scores in a cumulative sum approach.

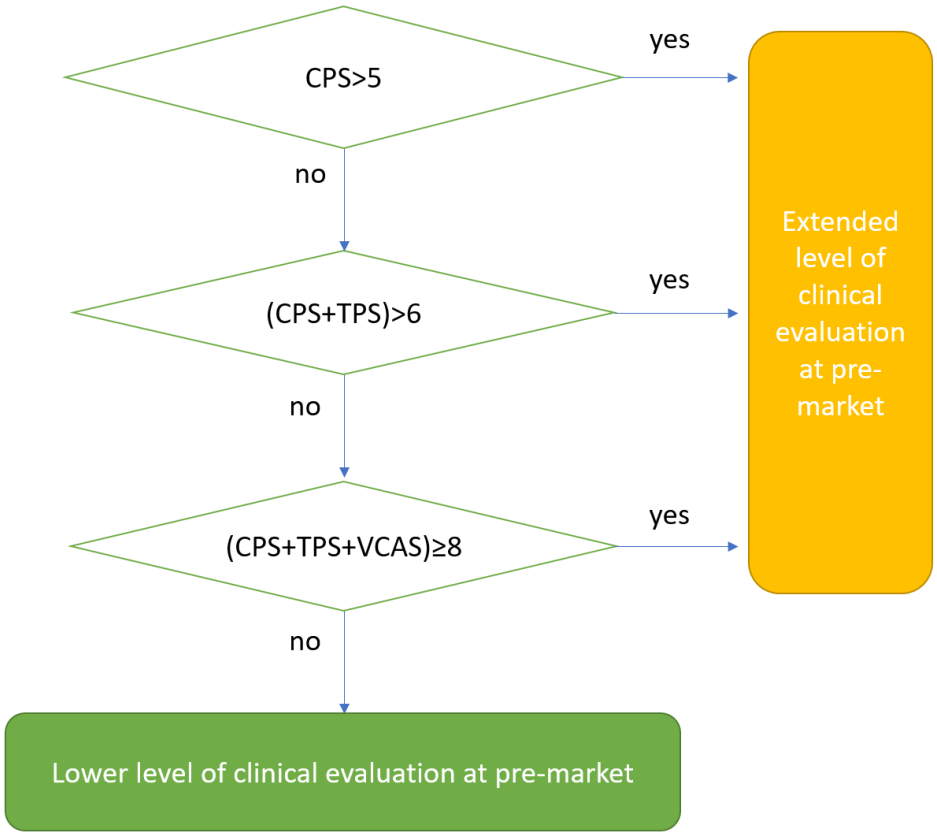

Figure 9. Flow chart accumulating different scoring systems

This score system (Figure 10) first evaluates the context in which the AI tool is claimed to be utilized, by requesting an extensive evaluation in case of use for diagnosis or treatment in critical conditions. Then, the score relevant to the level of validation performed on the system is added, and finally the score relevant to the valid clinical association, according to the supervised/unsupervised ML/DL model, is summed up. The final cumulative sum can vary between 4 and 12, thus defining a continuum between low-risk tools, only informing about low risk situations with full transparency, completely interpretable and explainable and with the possibility for comprehensive human oversight on the one hand, and autonomous systems deciding about treatments in critical conditions based on black-box algorithms without any possibility for human oversight on the other hand, with associated different evaluation requirements for release of such AI tools, thus defining the depth or level of evidence to be reached at each stage of the lifecycle.

In this way, low-risk tools only informing about low risk situations, with only internal validation and with easily verified transparency, will be associated to a lower level of pre-market clinical evaluation; however, if their application is in serious conditions or to drive the decision, the requirement for a more extensive pre-market evaluation will arise, unless compensated by a higher level of validation or by the use of a more transparent and explainable approach. Still HCP's will need to consider, even for low-risk tools, if

their usability and added time and administrative burden are acceptable before considering to use the tool in their practice.

For all the AI tools, once on the market, an adequate algorithmic vigilance process should be implemented and performed [89] in view of inevitable drift and the difficulty of accessing real clinical risk and benefit outside of real-life use.

4. Requirement Matrix

Depending on the risk score flowchart and the resulting overall score the level/depth of the generic requirements for clinical evaluation pre-release are defined:

- Extended level of clinical evaluation pre-release
- Limited level of clinical evaluation pre-release

with specific content at the pre- and post-release phases, keeping in mind the proposed 10 principles of Good Machine Learning Practices.

Requirement categories

- **Pre-release**
  - Limited level of evidence
    - Score 4 A
    - Score 5-6 B
    - Score 7 C
  - Extended level of evidence
    - Score 8 D
    - Score 9-10 E
    - Score 11-12 F
- **Post-release**
  - Limited level of evidence
    - Score 4 A
    - Score 5-6 B
    - Score 7 C
  - Extended level of evidence
    - Score 8 D
    - Score 9-10 E
    - Score 11-12 F

Rather than focusing only on the implementation phase, the relevant requirements can be described pertaining to all the AI life cycle stages as adapted from NIST [81]:

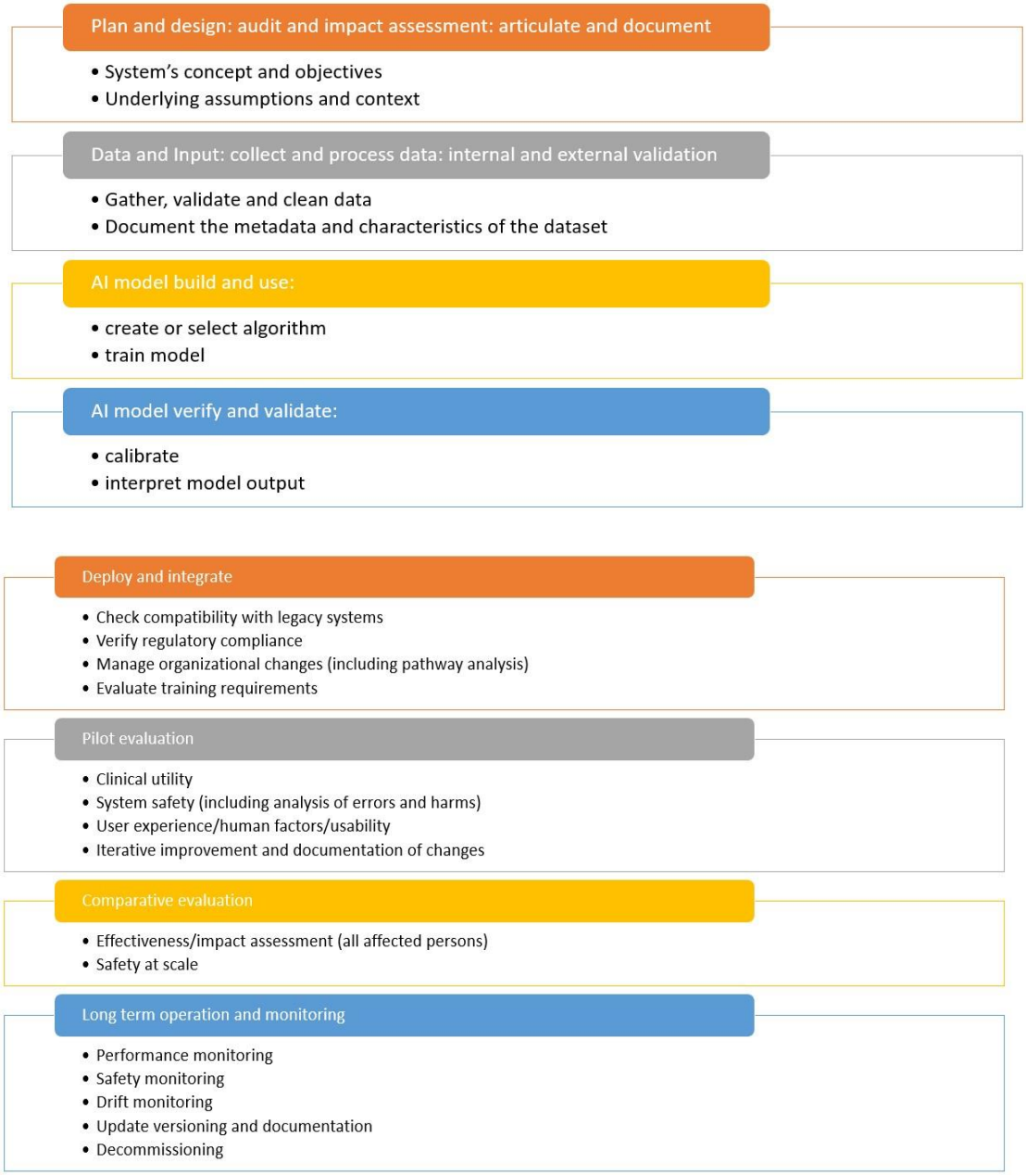

Figure 10. Requirements across AI life cycle

While some of these will always be necessary pre-release the extent/depth to which they are applied in the pre- or post-release phase is determined by the risk categories; some are less/never relevant in the post-release phase.

Each of these can always be applied in more or less depth and/or frequency (in the post-release phase) and this needs to be proportional to the benefit/risk ratio and is therefore specific to each use case.

Expert panels could play a crucial role in establishing interpretation and guidance of rules for clinical evaluation but need to include all stakeholders, HCP's as well as patients/citizens and might need to be present for other categories than only III).

Table 4. Requirement matrix (+ : required to be performed; - : not required to be performed)

Phase	AI life-cycle stages	Sub-stages	Requirement categories		
			A B C	D E F	comment
Pre release	Plan and design: audit and impact assessment: articulate and document	System's concept and objectives	+	+	
		Underlying assumptions and context	+	+	
	Data and Input: collect and process data: internal and external validation	Gather, validate and clean data	+	+	
		Document the metadata and characteristics of the datasets	+	+	
	AI model build and use	create or select algorithm	+	+	
		train model	+	+	
	AI model verify and validate	calibrate	+	+	
		interpret model output	+	+	
	Deploy and integrate	Check compatibility with legacy systems	+	+	
		Verify regulatory compliance	+	+	
		Manage organizational changes (including pathway analysis)	-	+	
		Evaluate training requirements	-	+	
	Pilot evaluation	Clinical utility	+	+	
		System safety	+	+	

		(including analysis of errors and harms)			
		User experience/human factors/usability	-	+	
		Iterative improvement and documentation of changes	-	+	
	Comparative evaluation	Effectiveness/impact assessment (all affected persons)	-	+	
		Safety at scale	-	+	
	Long term operation and monitoring	Performance monitoring	-	-	
		Safety monitoring	-	-	
		Drift monitoring	-	-	
		Update versioning and documentation	-	-	
		Decommissioning	-	+	
Post release	Plan and design: audit and impact assessment: articulate and document	System's concept and objectives	+	+	Drift
		Underlying assumptions and context	+	+	Drift
	Data and Input: collect and process data: internal and external validation	Gather, validate and clean data	+	+	Depending on change
		Document the metadata and characteristics of the datasets	-	-	Unless changed
	AI model build and use	create or select algorithm	-	-	Unless changed
		train model	-	-	
	AI model verify and	calibrate	-	-	

	validate				
		interpret model output	-	-	
	Deploy and integrate	Check compatibility with legacy systems	+	+	
		Verify regulatory compliance	+	+	
		Manage organizational changes (including pathway analysis)	+	+	
		Evaluate training requirements	+	+	
	Pilot evaluation	Clinical utility	+	+	
		System safety (including analysis of errors and harms)	+	+	
		User experience/human factors/usability	+	+	
		Iterative improvement and documentation of changes	+	+	
	Comparative evaluation	Effectiveness/impact assessment (all affected persons)	+	+	
		Safety at scale	+	+	
	Long term operation and monitoring	Performance monitoring	+	+	
		Safety monitoring	+	+	
		Drift monitoring	+	+	
		Update versioning and documentation	+	+	
		Decommissioning	+	+	

In case of changes to the tool itself (either self-learning or organised) or to the intended use, user, environment, the same risk assessment scoring system could be used and depending on the score change, the (supplementary) evidence/data become required.

5. Practical implementation of Clinical Evaluation in different stages of AI MDSW life cycle: Examples adapted from reference [36].

5.1 Plan and design: audit and impact assessment: articulate and document

- Has the manufacturer created a list of all roles that are directly or indirectly concerned with AI and defined the AI – related skills for each role? Including end-users: HCP's, patients, citizens?
- Has the manufacturer determined for which medical purpose (prevention, diagnosis, therapy, monitoring, predictions) the system is to be used and for which parts of the intended use an AI is to be used?
- Has the manufacturer characterized the individuals to be diagnosed, treated or monitored with the medical device? Does this characterization includes indications, contraindications and associated diseases?
- For which individuals the device is not to be used?
- Has the manufacturer specified from where (human, environment, existing sets) the data, used to train and test the device, originate?
- Has the manufacturer characterized the intended users, e.g.
    - using demographic features (age, gender),
    - regarding the training and experience in medical domains,
    - regarding technical knowledge,
- using physical and mental limitations, linguistic skills and cultural background?
- Has the manufacturer characterised the intended use environment both physical (imaging equipment, IT requirements) and human (also with regard to the social environment, influenced by stress, shift work, frequently changing colleagues, etc.)?
- Has the manufacturer defined the outcomes which can be impacted by the device? Are they measurable, quantifiable, easily available in routine clinical care (avoiding admin overload when gathered in post-release phase)?
- Have the stakeholder requirements been identified by the manufacturer and translated accordingly into the performance specifications?
- Has the manufacturer listed alternative methods to AI and evaluated them with regard to benefit, safety and performance?
- Has the manufacturer justified why AI is superior to conventional methods and thus justifies the associated risks and determined the clinical benefit of using the AI MDSW with respect to meaningful, measurable, patient-relevant clinical outcomes?
- Has the manufacturer drawn up a list of risks specifically arising from the use of AI techniques?

- Has the manufacturer analysed the risks that arise when users other than the specified users use the product?
- Has the manufacturer analysed the risks arising through use in an environment different than that specified?
- Has the manufacturer analysed the risks posed by inputs that do not meet the specified formats and/or have not been generated according to the specified prerequisites?
- Has the manufacturer analysed the risks that arise if the outputs do not meet the specified quality criteria?
- Has the manufacturer assessed the risks if the system is used in a different patient population than specified?
- Has the manufacturer derived the quantitative quality criteria based on the state of the art?
    - Peer-reviewed relevant literature
- Has the manufacturer defined operational limits within which the AI system may operate?
- Has the manufacturer defined how to ensure that these operational limits are not exceeded?
- Has the manufacturer assessed the risks if the system is not available?
- Has the manufacturer derived quantitative quality criteria or requirements for the software or/and the algorithm from the intended use in a comprehensible way?

5.2 Data and Input: collect and process data: internal and external validation

- Is the training data set representative of the actual patient population? Has the manufacturer assessed the consequences if the system provides socially unacceptable outputs (e.g. discriminatory)?
- Has the manufacturer justified where it collects test data and why it is representative of the target population? Where appropriate, has it compared these with data from the Federal Statistical

Office, scientific publications (prospective – retrospective studies), data from curated databases/registries/reference databases, data from equivalent devices?

- Has the manufacturer listed and discussed factors that could cause a bias of the validation and test data?
- Has the manufacturer analysed what influences the type and location of data collection has on the data?
- Has the manufacturer established a procedure to anonymise or pseudonymise data before training and testing?
- Has the manufacturer investigated and ruled out the possibility of label leakage?
- In the case of supervised learning, did the manufacturer derive the labels from the intended use for which the training data is understood and justify this choice?
- In the case of supervised learning, did the manufacturer specify a procedure for labelling if no labels were yet present in the data?
- Does this procedure specify quantitative/qualitative classification criteria for labeling? Has the manufacturer justified the choice of these criteria?
- Does this procedure specify the requirements for the number, training and competence of the persons responsible for labeling?
- Has the manufacturer set a procedure describing the (pre-)processing of the data?
- Does this procedure describe the individual processing steps such as conversions, transformations, aggregations, normalisation, format conversions, calculation of features and conversion of numerical data into categories (augmentation)?
- Does this procedure specify how values determined with different measurement methods are detected and processed?
- Does this procedure specify how missing values, outliers and unusable data within data sets are detected and processed? Has the manufacturer justified this specification?
- Has the manufacturer described the collected data using descriptive statistics? See also Dataset Nutrition Label (https://ahmedhosny.github.io/datanutrition/)
- Has the manufacturer characterised the inclusion and exclusion criteria of data using relevant attributes?
- Has the manufacturer specified technical inclusion and exclusion criteria for data?
- Has the manufacturer described the procedure to ensure that records that do not meet the inclusion criteria or are to be excluded are in fact excluded?

5.3 AI model build and use

- Is it documented in the product file which goal the machine learning procedures pursue?

- Did the manufacturer justify the final choice of model on the basis of the quality criteria and the intended use, and in particular explain when simpler and more interpretable models were not used?
- With Continuous Learning Systems, has the manufacturer considered the option of resetting the system to a known status?
- With Continuous Learning Systems, has the manufacturer shown quantitatively why the risk-benefit analysis is better than for non-continuously learning systems?
- Has the manufacturer specified the number of records and given a justification as to why this is sufficient?
- Has the manufacturer justified the selection of the features considered during training?
- Has the manufacturer described the interdependence of the features, especially in the case of tabular data?
- Has the manufacturer documented and justified the ratio in which it divides the data into training, validation and test data?
- Has the manufacturer documented the stratification used to divide the data into training, validation and test data?

5.4 AI model verify and validate

- Has the manufacturer specified the data interfaces, including the formats and, in the case of images, their specific properties (size, resolution, colour coding)?
- Has the manufacturer determined, documented and justified the quality metrics based on the intended use for which he wants to optimise the model?
- Has the manufacturer trained and compared several model types (including simpler and interpretable models), where appropriate?
- Has the manufacturer considered the following quantitative quality criteria or requirements
    - for classification problems: accuracy (resulting from trueness and precision) (mean or balanced accuracy), positive predictive value (precision), specificity and sensitivity, positive predictive value, negative predictive value, number needed to treat (average number of patients that need to be diagnosed/ treated in order to have an impact on one person), number needed to harm (number of patients that need to be diagnosed/ treated

in order have an adverse effect on one patient), positive likelihood ratio, negative likelihood ratio, odds ratio, confidence intervals;
    - for regression problems: mean absolute error and mean square error?
    - limit of detection,
    - limit of quantitation,
    - analytical specificity,
    - linearity,
    - cut-off value(s),
    - measuring interval (range),
- Qualitatively
    - availability,
    - confidentiality,
    - integrity,
    - reliability,
    - generalisability,
    - expected data rate or quality,
    - absence of unacceptable cybersecurity vulnerabilities
    - human factors engineering.
- Has the manufacturer specified the expected value ranges of the outputs?
- Has the manufacturer specified the requirements regarding repeatability and reproducibility of requirements?
- Has the manufacturer documented the quality metrics for the different models, e.g. for a binary classification, with the help of a confusion table?
- Has the manufacturer assessed and documented the quality metrics for the different models not only globally, but also separately for different features, if applicable?
- Has the manufacturer examined the data sets that predicted particularly well and those that predicted particularly poorly?
- Has the manufacturer examined the data sets for which the model decision is particularly safe and particularly unsafe?
- For tabular data in particular, has the manufacturer considered displaying, for individual data sets, the features that particularly drove the model to make the decision (Explainable AI)?
- For tabular data in particular, has the manufacturer considered evaluating how and to what extent individual features would have to change for the model to arrive at a different prediction?
- For tabular data in particular, has the manufacturer considered analysing / visualizing the dependence (strength, direction) of the predictions on the feature values?
- Has the manufacturer considered synthesizing data sets that particularly activate the model?
- In case the demonstration of conformity with GSPRs based on clinical data is not deemed appropriate (MDR Article 61 (10)), has the manufacturer duly substantiated in the technical

documentation why it is adequate to demonstrate conformity based on the results of non-clinical testing methods alone, including PERFORMANCE EVALUATION, bench testing and preclinical evaluation, and USABILITY assessment

5.5 Deploy and integrate

- Has the manufacturer determined the run-time environment of the product in terms of hardware (screen size, screen resolution, memory, network connection, etc.) and software (e.g. operating system, browser, run-time environments such as Java Run-time Environment or. .NET)?
- Has the manufacturer specified the input data requirements?
- Has the manufacturer identified the cybersecurity risks applicable to the AI, such as poisoning attacks, evasion attacks or model extraction etc.?
- Do the instructions for use explicitly state the patients / data / use cases for which the product may not be used?
- Do the instructions for use document the requirements for the input data (including formats, resolutions, range of values, etc.)?
- Do the instructions for use specify the intended primary and secondary users according to the intended use?
- Has the manufacturer determined whether an instructions for use and training materials are required?
- Do the instructions for use identify the version of the product with sufficient precision?
- Do the instructions for use describe the intended use of the product including the expected medical benefit?
- Do the instructions for use identify the intended patient population on the basis of indications, contraindications and – where relevant – other parameters such as age, gender, concomitant diseases or availability of information?
- Do the instructions for use describe what other prerequisites the product assumes (e.g. runtime environment, usage environment)?
- Do the instructions for use describe the residual risks?
- If useful: Do the instructions for use specify the data with which the model was trained?
- If useful: Do the instructions for use describe the model or the algorithms?
- If useful: Do the instructions for use specify the quality criteria?
- Has the manufacturer identified legacy systems and tested compatibility and/or transfer of content?
- Has the manufacturer provided the correct arguments why a prospective/retrospective analysis is more appropriate to support compliance with the General Safety and Performance Requirements

- Has the manufacturer evaluated the resistances to introduction of the product in the intended use environment and suggested means to counter them?
- Should the system be limited to populations/equipment with certain characteristics?
- Are there any circumstances where the system should not be used? Is the Quality system based on ISO 13485 Quality System Management and adhering to additional local requirements in regulations?
- Is the software development life cycle process based on IEC 62304 Software Life Cycle Process?
- Is a clinical investigation according to ISO 14155 required and, if so, performed?
- Is a risk management according to ISO 14971 performed, including AI specific risk such as bias?

5.6 Pilot Evaluation

Has the manufacturer specified what the user interface must display if the requirements are not met in order to operate the system safely (e.g. inputs not valid or not expected)?

- Has the manufacturer determined whether a quality of output needs to be provided to the user?
- If so, how is the quality indicated to the user?
- As part of the usability validation, does the manufacturer assess whether the users understand the instructions for use?
- As part of the usability validation, does the manufacturer assess whether users blindly trust the product or check the results?
- As part of the usability validation, does the manufacturer assess whether the users correctly recognize and understand the results?
- Has the manufacturer specified how fast the system must generate the outputs?
- Could a clinician correct an error if it occurred?
- Is the amount of human oversight proportional to the clinical risk posed by the AI system?
- Is it clear who is accountable for decisions made/influenced by the AI system?
- Is it clear which version of the system (and each individual component) is being used?
- Is there a plan to re-build the AI system should errors or biases be identified?
- Can a clinician understand the AI system output?
- Can a clinician understand how the AI system produced the output?
- Could a suitably trained human observer spot an error (and intervene)?

5.7 Comparative evaluation

- As part of the clinical evaluation, does the manufacturer assess whether the promised medical benefit corresponds to the state of the art?
- Has the manufacturer specified how the system will behave if the inputs do not meet the specified requirements?

- What requirements must be met in order to be able to detect misconduct, e.g. by means of self-tests?
- Have all individuals, possibly affected by the use of the system, been identified?
- Has the possible impact on these individuals been defined?

5.8 Long-term operation and monitoring

- Has the manufacturer prepared a Post-Market Surveillance (PMS) Plan?
- Has the manufacturer specified in this PMS plan the data he intends to collect and evaluate?
- Has the manufacturer specified in the PMS plan at which quality criteria and thresholds it considers action necessary, in particular a reassessment of the risk-benefit balance?
- When setting these thresholds, has the manufacturer analysed which feedback loops may influence the thresholds themselves?
- When setting these thresholds, has the manufacturer analysed which self-fulfilling prophecies may influence the thresholds themselves?
- Has the manufacturer described in the PMS plan how it collects and analyses what information on adverse medical effects?
- Does the manufacturer assess in the clinical evaluation whether the promised medical benefit is achieved with the given quality parameters?
- Has the manufacturer established that the clinical evaluation lists alternative methods, technologies or procedures, including their risks and benefits?
- Has the manufacturer described in the PMS plan how and which information on (adverse) behavioural changes or (predictable) misuse is collected and how this information is assessed?
- Has the manufacturer described in the PMS plan how it collects and assesses information on additional "adverse effects"?
- Has the manufacturer described in the PMS plan how and which information is collected to assess whether the data in the field is consistent to the expected data or training data?
- Has the manufacturer described in the PMS plan how and how often it will collect information on whether the product is still state-of-the-art?
- Has the manufacturer described in the PMS plan how and how often it will collect information on whether the ground truth or gold standard is still current?
- Has the manufacturer described in the PMS plan how and how often it verifies that changes are compliant with the regulatory requirements?

6. Summary and conclusions

This approach to evaluate the clinical impact of AI tools for clinical decision support requires further validation by the stakeholders, practical implementation (templates and interactive tools) and has several potential drawbacks.

The risk-benefit assessment is complex, and many factors contribute to the ultimate result which are often not (fully) quantifiable and subjective. Reducing them to a simple scoring system could both overestimate and underestimate the true risk-benefit ratio but has the advantage of clarity and simplicity. Reduction of risk to a zero level is both unattainable and risks to leave citizens and patients not having the benefit of tools that while a certain risk could contribute significantly to an improved desired outcome. Only the interaction between an informed patient and the HCP can lead to real co-decision making, balancing the risks and benefits of every choice to be made.

The use of the requirement matrix, based on the risk-benefit assessment, to define the scope and depth of the various obligations in the different life-cycle stages of the AI tool, could amplify errors in the assessment of the risk-benefit ratio, so they remain guiding principles which should be evaluated in each specific case.

The practical implementation using a set of questions to be answered by the manufacturer to verify the adequate clinical evaluation of the AI tool has the risk of becoming a “tick the box” exercise without proper comprehensive evaluation of all the possible impacts of the tool in testing and real-life circumstances.

Therefore, an alternative approach could be the direction the FDA is taking for the moment with a more case by case evaluation and the creation of a jurisprudence, acknowledging the needed flexibility of AI/ML-SaMD and the network interaction between devices and system level considerations. This type of approach would acknowledge the extremely complex conditions with interlinked data structures and origins and a shift of the centre of gravity for management from the hospital to the home. People seek optimal health not just cure and care and want to be able to self-manage their lives when desired, avoiding conservative and protectionist mechanisms, sometimes embedded in the present system. It means to keep the proper balance between benefits and harms due to irresponsible innovations versus over regulated conservatism. Regulatory science tries to keep this balance but still needs to be expanded and further implemented. The present EU legislation focusses mainly on “product legislation” with an added layer of horizontal AI regulation but missing the aspect of network effects between digital systems, digital

twins, medical devices and non-devices, interoperable patient experiences and clinical decision support tools.

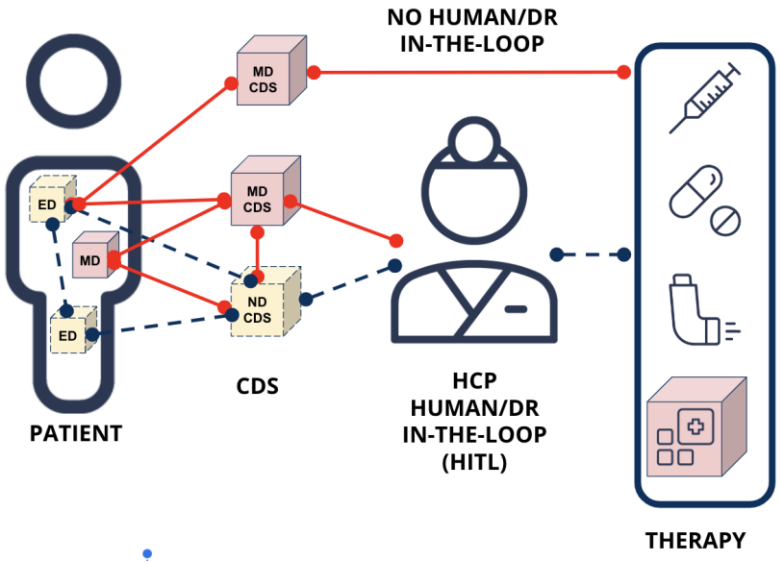

Figure 11. AI MDSW with Human in the loop (HITL)

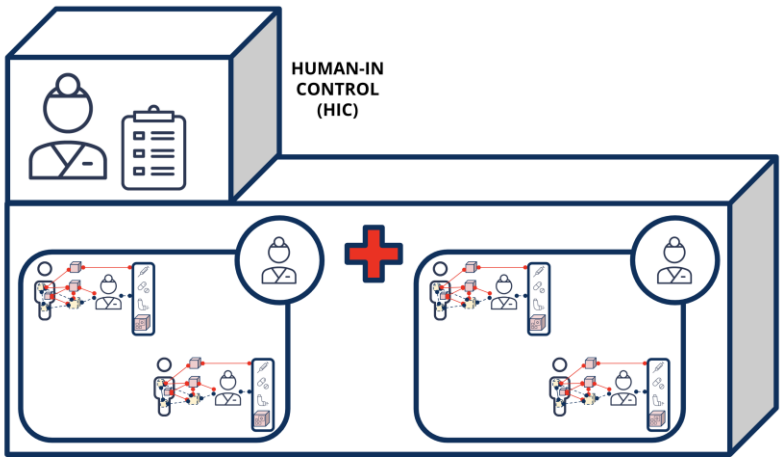

Figure 12. AI MDSW with Human in control (HIC)

In addition, health care providers want to move into more digitally supported systems in which the characteristics and environment of the individual is taken more into account, requiring more elaborate decision support tools to do so, but also considering the feedback and experiences of individual end-users (HCPs, patients and healthy individuals).

The FDA proposes a new ACP/PCCP/SPS model to create a more agile approach to AI evaluation with predetermined change control plans allowing a continuous cycle of monitoring, improvement, approval and delivery.

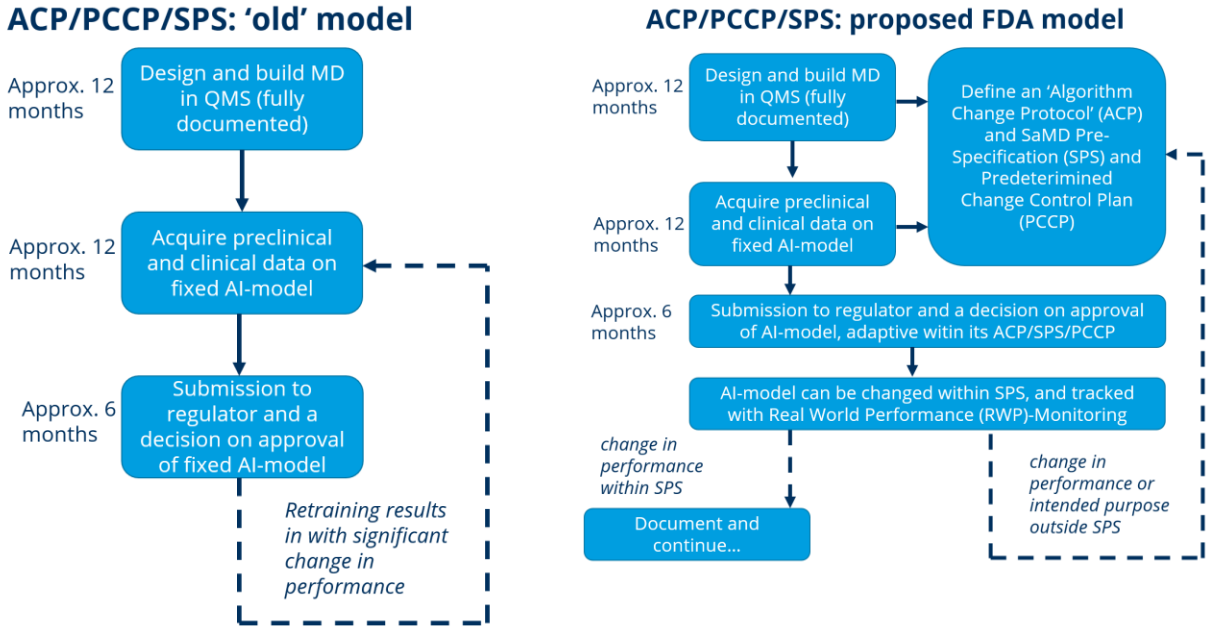

Figure 13. FDA model for AI evaluation

In contrast some interpretation of the present legal system in the EU would preclude any certification of non-static AI tools: "Static AI (AI that has learned and operates in a learned state) is in principle certifiable. Dynamic AI (AI that continues to learn in the field) is not certifiable in principle, as the system must be verified and validated (among other things, the functionality must be validated against the intended use)." And this also negates the existence of a continuum between pure static and fully adaptive ML tools which would necessitate a (near) automation of QMS processes and real-time technical dossier updates.

Adopting such an agile approach to clinical evaluation of AI clinical decision tool will require revision of the present (interpretation of the) regulation and much more emphasis on continued evaluation and re-evaluation in real time conditions with Human-in-Control (HIC) mechanisms, either In, On or Off (absent from) the Loop.

Notwithstanding these limitations and considerations, this paper is the end product of a long series of interactions and inputs by a broad field of stakeholders and experts from within and outside the CORE-MD consortium, the main strength of this project bringing them all together. Next steps will involve a Delphi-like validation of this proposal as well as a further practical implementation suggestion. As this is a complex and fast-moving field further interactions with stakeholders across the board will be required to ultimately deliver what the MDR legislation is intending: improved and safer MDSW bringing better care to the citizens who require it.

References

[1] Stapelfeldt R. Can there be a dumb Super intelligence? A critical look at Bostrom's notion of Sup. Academia Letters. Published online July 15, 2021. doi:10.20935/al2076

[2] OECD digital economy papers: OECD FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS.; 2022. www.oecd.ai/wips.

[3] Lii S. Ministero Della Salute Consiglio Superiore Di Sanità Sezione V "I Sistemi Di Intelligenza Artificiale Come Strumento Di Supporto Alla Diagnostica."; 2019.

[4] Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1). doi:10.1186/s12916-019-1466-7

[5] Van Smeden M, Heinze G, Van Calster B, Asselbergs FW, Vardas PE, Bruining N, de Jaegere P, Moore JH, Denaxas S, Boulesteix AL, Moons KGM. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. Eur Heart J. 2022 Aug 14;43(31):2921-2930. doi: 10.1093/eurheartj/ehac238. PMID: 35639667; PMCID: PMC9443991.

[6] Ozkan J. Thinking outside the black box: CardioPulse takes a look at some of the issues raised by machine learning and artificial intelligence. Eur Heart J. 2023;44(12):1007-1009. doi:10.1093/eurheartj/ehac790

[7] Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: A survey study from consumers' perspectives. BMC Med Inform Decis Mak. 2020;20(1). doi:10.1186/s12911-020-01191-1

[8] Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-38. doi:10.1038/s41591-021-01614-0

[9] De Nigris S., Craglia M., Nepelski D., et al. AI Watch : AI Uptake in Health and Healthcare, 2020.

[10] Lekadir Karim, Quaglio Gianluca, Tselioudis Garmendia Anna, Gallin Catherine, European Parliament. Directorate-General for Parliamentary Research Services. Artificial Intelligence in Healthcare : Applications, Risks, and Ethical and Societal Impacts.

[11] Matheny ME, Whicher D, Thadaney Israni S. Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. JAMA. 2020 Feb 11;323(6):509-510. doi: 10.1001/jama.2019.21579. PMID: 31845963.

[12] Clinical Investigation of Medical Devices for Human Subjects-Good Clinical Practice.; 2020. www.iso.org

[13] European Commission, Directorate-General for Communications Networks, Content and Technology, HLEG on AI (2022). Policy and investment recommendations for trustworthy AI, Publications Office of the European Union. https://data.europa.eu/doi/10.2759/465913

[14] European Commission, Directorate-General for Communications Networks, Content and Technology, HLEG on AI (2019), Ethics guidelines for trustworthy AI, Publications Office, 2019, https://data.europa.eu/doi/10.2759/346720

[15] WHO, Ethics and Governance of Artificial Intelligence for Health 2.; 2021. https://www.who.int/publications/i/item/9789240029200.

[16] Tschider, C. (2018), Regulating the IoT: Discrimination, Privacy, and Cybersecurity in the Artificial Intelligence Age (February 24, 2018). 96 DENV. U. L. REV. 87 (2018), Available at SSRN: http://dx.doi.org/10.2139/ssrn.3129557.

[17] Wachter, Sandra and Mittelstadt, Brent, A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI (October 5, 2018). Columbia Business Law Review, 2019(2), Available at SSRN: https://ssrn.com/abstract=3248829

[18]Simon Chesterman, Through a Glass, Darkly: Artificial Intelligence and the Problem of Opacity, The American Journal of Comparative Law, Volume 69, Issue 2, June 2021, Pages 271–294, https://doi.org/10.1093/ajcl/avab012.

[19] Doshi-Velez, Finale, and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017).

[20] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-e750. doi:10.1016/S2589-7500(21)00208-9

[21] Nicholson Price W. Big data and black-box medical algorithms. Sci Transl Med. 2018;10(471). doi:10.1126/scitranslmed.aao5333

[22] Pasquale FA. The Black Box Society: The Secret Algorithms That Control Money and Information. Vol Book Gallery. 96. (Francis King Carey School of Law U of M, ed.). Harvard University Press; 2015.

[23] Vasey B, Ursprung S, Beddoe B, et al. Association of Clinician Diagnostic Performance with Machine Learning-Based Decision Support Systems: A Systematic Review. JAMA Netw Open. 2021;4(3). doi:10.1001/jamanetworkopen.2021.1276

[24] Dey D, Arnaout R, Proceedings of the NHLBI Workshop on Artificial Intelligence in Cardiovascular Imaging: Translating AI to Patient Care. 2022.

[25] Gilbert S, Fenech M, Hirsch M, Upadhyay S, Biasiucci A, Starlinger J. Algorithm change protocols in the regulation of adaptive machine learning-based medical devices. J Med Internet Res. 2021;23(10). doi:10.2196/30545

[26] Regulatory Guidelines for Software Medical Devices-A Lifecycle Approach. 2019. Health Sciences Authority Singapore.

[27] Zhang J, Budhdeo S, William W, et al. Moving towards vertically integrated artificial intelligence development. NPJ Digit Med. 2022;5(1). doi:10.1038/s41746-022-00690-x

[28] Gerke S, Babic B, Evgeniou T, Cohen IG. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. NPJ Digit Med. 2020;3(1). doi:10.1038/s41746-020-0262-2

[29] Banerjee, I., Bhimireddy, A.R., Burns, J.L., Celi, L.A., Chen, L., Correa, R., Dullerud, N., Ghassemi, M., Huang, S., Kuo, P., Lungren, M.P., Palmer, L.J., Price, B.J., Purkayastha, S., Pyrros, A., Oakden-Rayner, L., Okechukwu, C., Seyyed-Kalantari, L., Trivedi, H., Wang, R., Zaiman, Z., Zhang, H., & Gichoya, J.W. (2021). Reading Race: AI Recognises Patient's Racial Identity In Medical Images. ArXiv, abs/2107.10356..

[30] Richardson JP, Smith C, Curtis S, et al. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med. 2021;4(1). doi:10.1038/s41746-021-00509-1

[31] Mittelstadt B. THE IMPACT OF ARTIFICIAL INTELLIGENCE ON THE DOCTOR-PATIENT RELATIONSHIP. Council of Europe. Human rights and Biomedicine; now CDBIO. 2021.
[32] European Commission, Joint Research Centre, Balahur, A., Jenet, A., Hupont Torres, I., et al., Data quality requirements for inclusive, non-biased and trustworthy AI : putting science into standards, Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2760/365479
[33] Association for Progressive communications (APC), Article 19, And Swedish International Development cooperation Agency (SIDA). Artificial Intelligence: Human Rights, Social Justice and Development. GLOBAL INFORMATION SOCIETY WATCH. 2019 Report. www.GISWatch.org
[34] Comandè G, Schneider G. Regulatory Challenges of Data Mining Practices: The Case of the Never-ending Lifecycles of "Health Data." Eur J Health Law. 2018;25(3):284-307. doi:10.1163/15718093-12520368
[35] Artificial Intelligence in Medical Devices Verifying and Validating AI-Based Medical Devices. TUV Sud report. 2022.
[36]Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. Ann Intern Med. 2019;170(1):W1-W33. doi:10.7326/M18-1377
[37] Messina D. Online Platforms, Profiling, and Artificial Intelligence: New Challenges for the GDPR and, in Particular, for the Informed and Unambiguous Data Subject's Consent. Media Laws. Rivista di diritto dei media 2/2019.
[38] On Artificial Intelligence-A European Approach to Excellence and Trust White Paper on Artificial Intelligence A European Approach to Excellence and Trust. https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf.
[39] Smuha NA. Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea. Philos Technol. 2021;34:91-104. doi:10.1007/s13347-020-00403-w
[40] AI Act EU. https://artificialintelligenceact.eu/
[41] Floridi, Luciano and Holweg, Matthias and Taddeo, Mariarosaria and Amaya Silva, Javier and Mökander, Jakob and Wen, Yuni, capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act (March 23, 2022). Available at SSRN: https://ssrn.com/abstract=4064091 or http://dx.doi.org/10.2139/ssrn.4064091.
[42] Cobbaert K. AI Act in Healthcare Impact for Medical Device Manufacturers and Clinicians. Internal document CORE-MD
[43] IMDRF. Machine Learning-Enabled Medical Devices-A Subset of Artificial Intelligence-Enabled Medical Devices: Key Terms and Definitions Authoring Group: IMDRF AIMD Working Group.; 2021.
[44] Software as a Medical Device (SAMD): Clinical Evaluation - Guidance for Industry and Food and Drug Administration Staff. 2017. https://www.fda.gov/MedicalDevices/InternationalPrograms/IMDRF/default.htm.
[45] Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. 2022. doi:10.6028/NIST.SP.1270

[46] Risk Management Framework for Information Systems and Organizations: 2018. NIST Special Publication 800-37 Revision 2 doi:10.6028/NIST.SP.800-37r2

[47] Liu Q, Naik K, Mehta N, et al. FDA-MCERSI Workshop on Application of Artificial Intelligence and Machine Learning for Precision Medicine Virtual Public Workshop SESSION 1: Background and Current Landscape.; 2023.

[48] General Wellness: Policy for Low Risk Devices Guidance for Industry and Food and Drug Administration Staff. 2019. https://www.regulations.gov.

[49] The Software Precertification (Pre-Cert) Pilot Program: Tailored Total Product Lifecycle Approaches and Key Findings. 2022. FDA.

[50] Lyell D, Coiera E, Chen J, Shah P, Magrabi F. How machine learning is embedded to support clinician decision making: An analysis of FDA-approved medical devices. BMJ Health Care Inform. 2021;28(1). doi:10.1136/bmjhci-2020-100301

[51] Roberts, H., Cowls, J., Morley, J. et al. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. AI & Soc 36, 59–77 (2021). https://doi.org/10.1007/s00146-020-00992-2

[52] Software as a Medical Device (SaMD), Definition and Classification. Health Canada Guidance document. 2019.

[53] Jorge Ricart, R., Van Roy, V., Rossetti, F. and Tangi, L., AI Watch. National strategies on Artificial Intelligence: A European perspective. 2022 edition, EUR 31083 EN, Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-76-52910-1, doi:10.2760/385851, JRC129123.

[54] Veronese A, Silveira A, Lopes AN, Lemos E. Artificial Intelligence, Digital Single Market and the Proposal of a Right to Fair and Reasonable Inferences: A Legal Issue between Ethics and Techniques. Vol 5.; 2019. https://eur-lex.europa.eu/legal-content/EN/TXT/

[55] Mongnani M.L. (2020), Virtues and Perils of Algorithmic Enforcement and Content Regulation in the EU - A Toolkit for a Balanced Algorithmic Copyright Enforcement, 11 Case W. Res. J.L. Tech. & Internet 1 (2020). https://scholarlycommons.law.case.edu/jolti/vol11/iss1/2.

[56] Spain Proposes to Pilot an Artificial Intelligence Sandbox to Implement Responsible AI with a Human-Centric Approach. Launch event June 2022, Brussels.

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

[58] Fraser A.G., Biasin E., Bijnens B., et al. (2023). Artificial Intelligence in Medical Device Software and High-Risk Medical Devices-a Review of Definitions, Expert Recommendations and Regulatory Initiatives. Accepted to be published on 'Expert Review of Medical Devices'.

[59] Mazzini, G., Scalzo, S. (2022), The Proposal for the Artificial Intelligence Act: Considerations around Some Key Concepts (May 2, 2022). Forthcoming in Università Ca' Foscari di Venezia - Dipartimento di Economia - Collana Centro Studi Giuridici - Wolters Kluver - CEDAM, Available at SSRN: http://dx.doi.org/10.2139/ssrn.4098809.

[60] Kasperbauer TJ. Conflicting roles for humans in learning health systems and AI-enabled healthcare. In: Journal of Evaluation in Clinical Practice. Vol 27. Blackwell Publishing Ltd; 2021:537-542. doi:10.1111/jep.13510

[61] Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: Phases of clinical research. JAMIA Open. 2020;3(3):326-331. doi:10.1093/JAMIAOPEN/OOAA033

[62] Brown P, Lloyd C, Souto-Otero M. The Prospects for Skills and Employment in an Age of Digital Disruption: A Cautionary Note. SKOPE Research Paper No. 127, November 2018. www.skope.ox.ac.uk

[63] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996 Jan 13;312(7023):71-2. doi: 10.1136/bmj.312.7023.71. PMID: 8555924; PMCID: PMC2349778.-

[64] Rozich JD, Howard RJ, Justeson JM, Macken PD, Lindsay ME, Resar RK. Standardization as a Mechanism to Improve Safety in Health Care. The Joint Commission Journal on Quality and Safety. 2004;30(1):5-14. doi: 10.1016/S1549-3741(04)30001-8

[65] Lillrank P, Groop PJ, Malmström TJ. Demand and supply-based operating modes--a framework for analyzing health care service production. Milbank Q. 2010 Dec;88(4):595-615. doi: 10.1111/j.1468-0009.2010.00613.x. PMID: 21166870; PMCID: PMC3037177

[66] Marelli L, Lievevrouw E, Van Hoyweghen I. Fit for purpose? The GDPR and the governance of European digital health. Policy Studies. 2020;41(5):447-467. doi:10.1080/01442872.2020.1724929

[67] European Commission, Consumers, Health, Agriculture and Food Executive Agency, Hansen, J., Wilson, P., Verhoeven, E., et al., Assessment of the EU Member States' rules on health data in the light of GDPR, Publications Office, 2021, https://data.europa.eu/doi/10.2818/546193.

[68] Verhoeven E., Kroneman M., Wilson P., et al. Assessment of the EU Member States' Rules on Health Data in the Light of GDPR : Country Fiches for All EU MS.

[69] Bayamlıoğlu, E., Transparency of Automated Decisions in the GDPR: An Attempt for Systemisation (January 7, 2018). Available at SSRN: https://ssrn.com/abstract=3097653 or http://dx.doi.org/10.2139/ssrn.3097653

[70] Gultekin Varkonyi G. Operability of the GDPR's Consent Rule in Intelligent Systems: Evaluating the Transparency Rule and the Right to Be Forgotten. Intelligent Environments 2019. Published online 2019. doi:10.3233/AISE190044

[71] Felzmann H, Villaronga EF, Lutz C, Tamò-Larrieux A. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data Soc. 2019;6(1). doi:10.1177/2053951719860542

[72] Kiseleva A, Kotzinos D, De Hert P. Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations. Front Artif Intell. 2022;5. doi:10.3389/frai.2022.879603

[73] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-933. doi:10.1038/s41591-022-01772-9

[74] Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021. U.S. Food and Drug Administration (FDA), Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA)

[75] Walsh I, Fishman D, Garcia-Gasulla D, et al. DOME: recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18(10):1122-1127. doi:10.1038/s41592-021-01205-4
[76] MDCG 2019-11 Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR. 2019.
[77] Medical Device Medical Device Coordination Group Document GUIDANCE NOTES FOR MANUFACTURERS OF CLASS I MEDICAL DEVICES. 2019.
[78] IMDRF Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations. IMDRF Software as a Medical Device (SaMD) Working Group 2014.
[79] NIST PRIVACY FRAMEWORK:; 2020. doi:10.6028/NIST.CSWP.01162020
[80] NIST AI Risk Management Framework Playbook-MEASURE.
[81] NIST AI Risk Management Framework Playbook-MANAGE.
[82] NIST AI Risk Management Framework Playbook-GOVERN.
[83] NIST Risk Management Framework Quick Start Guide. Roles and Responsibilities Crosswalk; 2021. https://nist.gov/rmf
[84] NIST Privacy Framework; 2020. https://www.nist.gov/privacy-framework.
[85] Newlands G, Lutz C, Fieseler C. Trading on the Unknown: Scenarios for the Future Value of Data. Law and Ethics of Human Rights. 2019;13(1):97-114. doi:10.1515/lehr-2019-0004
[86] Bernal-Delgado E, Cascini F, Dinis S, et al. Report on Architecture and Infrastructure Options to Support EHDS Services for Secondary Use of Data. TEHDAS WP7, Milestone 7.3 2023. www.tehdas.eu.
[87] Andanda P. Towards a Paradigm Shift in Governing Data Access and Related Intellectual Property Rights in Big Data and Health-Related Research. IIC International Review of Intellectual Property and Competition Law. 2019;50(9):1052-1081. doi:10.1007/s40319-019-00873-2
[88] Starkbaum J, Felt U. Negotiating the reuse of health-data: Research, Big Data, and the European General Data Protection Regulation. Big Data Soc. 2019;6(2). doi:10.1177/2053951719862594
[89] Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health. 2022;4(5):e384-e397. doi:10.1016/S2589-7500(22)00003-6
[90] IMDRF. Software as a Medical Device (SaMD): Clinical Evaluation. Software as a Medical Device Working Group 2017.
[91] REGULATION (EU) 2017/746 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU.

CORE-MD

CORE-MD, Coordinating Research and Evidence for Medical Devices, aims to translate expert scientific and clinical evidence on study designs for evaluating high-risk medical devices into advice for EU regulators.

For more information, visit:

www.core-md.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 965246.