npj | digital medicine

Perspective

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01459-8

CORE-MD clinical risk score for regulatory evaluation of artificial intelligence-based medical device software

Frank E. Rademakers[1] ✉, Elisabetta Biasin[2], Nico Bruining[3], Enrico G. Caiani[4,5], Rhodri H. Davies[6], Stephen H. Gilbert[7], Eric Kamenjasevic[8], Gearóid McGauran[9], Gearóid O'Connor[9], Jean-Baptiste Rouffet[10], Baptiste Vasey[11,12] & Alan G. Fraser[13,14]

The European CORE–MD consortium (Coordinating Research and Evidence for Medical Devices) proposes a score for medical devices incorporating artificial intelligence or machine learning algorithms. Its domains are summarised as valid clinical association, technical performance, and clinical performance. High scores indicate that extensive clinical investigations should be undertaken before regulatory approval, whereas lower scores indicate devices for which less pre-market clinical evaluation may be balanced by more post-market evidence.

Artificial Intelligence (AI) in all its forms is being used increasingly in healthcare and medicine by both caregivers and patients/citizens[1]. Until recently most applications were supporting diagnosis (analysing electrocardiograms, imaging, pathological specimens, skin lesions, and retinal pictures, etc.) but now AI methods are being employed in addition to estimating prognosis and predict the effects of treatment (personalisation); to detect and extract health data (using natural language processing); to assist in drug development; to monitor patients remotely; to communicate with patients (chatbots); and to personalise therapy through digital therapeutics and digiceuticals. Many more uses arrive each day[2]. Besides direct medical applications, the roles of AI are expanding into medical research, training (also via extended reality), public health, administration, and logistics. AI drives robotics and automates procedures and interventions, offering promise for more lean and efficient healthcare[3,4]. Early applications of large language models (LLM) and the possibilities of foundation models are being explored and used in clinical contexts, but they are not covered by this document.

Several individuals and institutions have warned of risks associated with the unbridled use of AI[5], or even suggested a temporary ban on further development. The European Union (EU) has developed horizontal laws relevant to healthcare and AI, such as the General Data Protection Regulation (GDPR), the Artificial Intelligence Act, and the European Health Data Space Regulation[6]. More importantly, medical device software (MDSW) that incorporates AI algorithms, whether it is standalone or integrated within a diagnostic or high-risk therapeutic device, requires conformity assessment for regulatory purposes before it is approved as a medical device for general use in clinical practice. In the EU the principles of device evaluation are prescribed by the Medical Device Regulation (MDR) and the In Vitro Diagnostic Medical Devices Regulation (IVDR), while worldwide most jurisdictions and many standards organisations and expert groups are also developing guidance[6].

European guidance for MDSW[7] is based on international recommendations but does not comprehensively describe the specific clinical evidence needed for medical AI software. Thus there is a need for recommendations for the regulatory evaluation of AI MDSW, that could balance its potential for major beneficial impacts in healthcare against the possibility for its misuse and negative effects on individuals and society.

In the EU, producing guidance is the responsibility of the Medical Device Coordination Group (MDCG) (Article 105 MDR), which is composed of representatives from national regulatory agencies and chaired by the European Commission. A call from the Horizon 2020 programme sought external expert advice on methodologies for the clinical investigation

[1]Emeritus Professor of Cardiology, KU Leuven, Leuven, Belgium. [2]Researcher in Law, Center for IT & IP Law (CiTiP), KU Leuven, Leuven, Belgium. [3]Department of Cardiology, Erasmus Medical Center, Thorax Center, Rotterdam, the Netherlands. [4]Department of Electronics, Information and Biomedical Engineering, Politecnico di Milano, Milan, Italy. [5]IRCCS Istituto Auxologico Italiano, Milan, Italy. [6]Institute of Cardiovascular Science, University College London, London, UK. [7]Professor for Medical Device Regulatory Science, Else Kröner Fresenius Center, for Digital Health, TUD Dresden University of Technology, Dresden, Germany. [8]Doctoral researcher in Law and Ethics, Center for IT & IP Law (CiTiP), KU Leuven, Leuven, Belgium. [9]Medical Officer, Medical Devices, Health Products Regulatory Authority, Dublin, Ireland. [10]Policy Advisor, European Affairs, European Federation of National Societies of Orthopaedics and Traumatology, Rolle, Switzerland. [11]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK. [12]Department of Surgery, Geneva University Hospital, Geneva, Switzerland. [13]Consultant Cardiologist, University Hospital of Wales, and Emeritus Professor of Cardiology, School of Medicine, Cardiff University, Heath Park, Cardiff, UK. [14]Cardiovascular Imaging and Dynamics, KU Leuven, Leuven, Belgium. ✉e-mail: frank.rademakers@kuleuven.be

of high-risk medical devices including those incorporating AI. The CORE–MD project (Coordinating Research and Evidence for Medical Devices) established a task force for that specific objective – namely to outline methodological principles for the clinical evaluation of AI MDSW during its full life cycle, applying a risk-benefit approach and focusing on pre- and post-release phases from both regulatory and end-user perspectives[8]. This article presents its final recommendations.

Methods

Membership

The task force was led by KU Leuven and composed of CORE–MD consortium members and others invited because of their complementary expertise. Backgrounds and relevant experience encompassed clinicians who have used AI to analyse medical images and other data types, clinicians qualified in computer science or as authors of relevant expert consensus statements, biomedical, electronics and informatics engineers, specialists in medical technology and regulatory science, lawyers expert in EU legislation and ethical considerations, and doctors from EU national regulatory agencies for medical devices. Manufacturers were not primary members of the CORE–MD consortium but their trade associations were represented on its international advisory board and so for this task, advisers were included because of their participation in international standards-setting bodies.

Review of existing guidance

A comprehensive analysis was undertaken of definitions, recommendations, and standards relating to the use of AI in healthcare and medical devices, published by national, European, and global organisations. The resulting publication includes details of the search strategies that were used[6]. It was concluded that the level of clinical evidence should be determined according to each application and should consider factors that contribute to risk, including accountability, transparency, and interpretability. Some principles are summarised below as the rationale for developing a risk score to guide proportionate clinical evaluation of AI MDSW.

Delphi consensus

After a series of strategy meetings among the members of the task force, the first version of the report was drafted in October 2022, and a two-stage Delphi process was organised throughout 2023. The meetings were held online; 33 clinical experts participated in the first session and 26 in the second one. Round 1 consisted of 11 voting statements and one free-text question; each was introduced briefly by the task leader, before independent voting. The threshold for statements to be adopted was 70% of positive responses. During Round 2, six statements that did not achieve 70% positive responses at the first vote, were revised and resubmitted to the experts. Participants were also invited to comment on the draft report. All statements that achieved consensus were integrated into this recommendation.

Consultations with regulators and notified bodies

The draft proposal was circulated with an explanatory note to members of the Clinical Investigation and Evaluation (CIE) and New Technologies (NT) Working Groups of the MDCG of the European Commission. Its key elements were presented by the task leader and the scientific coordinator of CORE–MD at meetings of CIE in April and November 2023, and at NT in June and December 2023. In parallel, the updated CORE–MD recommendations were sent to regulators and notified bodies, members of the consortium, and a team in the Joint Research Centre (JRC) of the European Commission (in Ispra, Italy) that is studying AI in medical technology[9]. Finally, the proposals were presented to a meeting of the CORE–MD Advisory Board, and in a discussion with leadership of the International Medical Device Regulators Forum (IMDRF) Working Group on AI medical devices. Account was taken of all comments received.

Details of this methodology are provided in Supplementary Information Section 2.

Considerations for regulating AI medical devices

Quality and transparency of clinical decisions

Healthcare providers apply evidence-based guidelines to optimise approaches to specific medical problems[10]. Common approaches can help practitioners improve patient-relevant outcomes, maximising the health of individuals and the population[11]. Deviation from guidelines may be warranted on the basis of patients' unique backgrounds, needs or expectations, in which case healthcare professionals (HCP) should be able to justify their decisions to the people affected. AI MDSW can be a powerful support tool to minimise unwanted variation, which is inherent in human judgements and decision-making[12]. For it to earn the trust of end-users (citizens, patients, and HCPs) the AI MDSW must have undergone appropriate clinical evaluation and be compliant with the relevant MDR requirements. In many circumstances, the information provided can then enable real informed co-decision-making between the patient and HCP, whether for diagnostic or therapeutic options.

Human oversight

The autonomy of AI systems and the degree of possible human supervision vary greatly, so most commentators stress the need to integrate AI tools into existing workflow, creating a 'Human–AI team'[13]. Interpretation and oversight can become difficult or even impossible when AI systems perform as 'black boxes'[14] with their logic remaining obscure even when explainability methods, which often remain inadequate, are applied[15–23]. Oversight may be less effective for less experienced users who paradoxically might benefit the most from such tools[24]. It is exactly for this reason we have recommended transparency of all input data and continued evaluation of the diagnostic performance of such algorithms.

On the other hand, providing real-time human oversight might reduce safety and decrease the performance of an AI tool[25] whose capabilities exceed the human and the human–AI team in terms of speed (faster reaction times), performance (more accurate and precise), being less prone to errors and more consistent. It remains, however, the sole responsibility of the caregiver, together with the patient, to co-decide about preventive, diagnostic or therapeutic measures using AI tools, while taking into account the patient's values, social and lifestyle factors, culture and accessible resources.

Some AI tools are available as apps to be used by citizens and patients without any involvement of HCPs, in which case oversight depends on the end-user[26]. Human oversight can be very effective and appropriate in many circumstances, but that should not be used to shift responsibility and accountability for the output of AI MDSW from the manufacturer solely to the supervising human. Decisions mostly depend on the context of use and are made by clinical teams of HCPs so it would be more appropriate to consider liability at the level of the manufacturer and the organisation using the MDSW. As with all medical tools, AI MDSW should be evaluated in the intended population for the specific purpose with appropriate clinical investigations before implementation. It is the goal of this article to provide a practical guide as to how and in what phase of the AI life cycle such investigations should be performed.

On-market adaptive approaches

The implementation of the objectives of personalised medicine may be supported by AI tools that adapt to specific use settings and patients' characteristics, after they have been placed on the market, by using a learning approach to adjust their parameters with continuous or intermittent implementation of changes. The high complexity of the post-release phase of AI MDSW makes Algorithm Change Protocols challenging[27]. Any drift in the intended use of an AI algorithm or in the target population where it is applied, perhaps because of evolving clinical practice, could change its risk and performance metrics. Additional data need to be collected and used to adapt the algorithm, which necessitates continuous evaluation after its release. An agile approach to the development, testing, and validation of AI tools[28,29] preferably with a system view rather than a pure device focus[30–32], should be facilitated by regulatory standards.

Similarity with clinical judgement

To better understand how an AI decision-support tool could be positioned in the clinical workflow, the tool can be compared to a clinical colleague from whom one receives advice before deciding on a diagnostic or therapeutic action. No one is infallible but a clinical colleague is trusted because of their verified licence, education, training, competence, ethical principles, and experience. An AI MDSW developer should provide similar proof, with documentation of safety and performance and verification by a notified body, as required in the EU by the MDR (and by the AI Act), and with clinical evaluation showing a positive balance between predefined benefits and any associated risks. That will not guarantee that the MDSW will function without any errors, but it should support improved outcomes at a reasonable cost compared to other methods, leaving the final co-decision to the patient and the HCP.

Results and recommendations

Process endorsed

Almost all experts (90%) consulted in the Delphi process supported the risk-benefit-based approach for evaluating AI medical devices, and 80% supported the concept of a scoring system to guide requirements for clinical evidence (see Supplementary Information Section 2). Most (88%) also concurred with the recommendation that low-risk AI medical devices showing a clear benefit could be brought to the market with graded evidence and formal requirements for post-market follow-up. There was no consensus among the consortium members to incorporate in this proposed text the alternative approach that has been adopted by some regulators [such as the Food and Drug Administration (FDA)] of certifying software companies on the basis of their quality-control systems and adapting the requirements for specific MDSW release depending on such certification. Such a mechanism could make it very difficult, if not impossible, for small and medium-sized enterprises and academic institutions to comply with these requirements.

Risk-based approach

Transparency is key in all healthcare interactions. When using AI MDSW, the HCP should be completely open to the patient about its use and about inherent benefits and disadvantages including explainability or lack thereof, the technical and clinical evidence supporting its use, any alternatives, and the consequences of non-use. For this, of course, the HCP needs access to the evidence supporting the claims for the AI MDSW for its defined purpose.

The balance between positive outcomes and possible safety risks or side effects needs to be considered at both the individual and the societal level. When demonstrating clinical benefit and/or improved efficiency in workflow, individual and societal human rights might conflict to a certain degree, so ethical considerations are paramount[33–36]. Medical device legislation requires compliance with safety and performance requirements, with a positive benefit-risk balance taking into account the acknowledged state of the art[37]. When evaluating a new AI tool in the context of the present state of the art, the tool must demonstrate an improved benefit-risk ratio, while keeping the absolute risk as low as possible. Implementing a new tool involves managing both an operational and a cultural change, which may be difficult for the end-users to accept, so a comparison of risks is crucial for reaching a decision.

Manufacturers are required to justify why the clinical evidence for their AI tool, which they provide, is appropriate and conforms with standards. European guidance, in line with recommendations from the IMDRF, considers three aspects to be crucial for the safe and effective use of MDSW (and by implication AI): valid clinical association, technical performance, and clinical performance[7].

Defining risk

Risk is a composite of the probability of an event and its severity. Specific challenges for managing risks of AI tools, at every stage in their life cycle, may include:

(1) difficulty in defining and measuring negative impact or magnitude of harm;
(2) tolerance of risks due to societal acceptance and preferences;
(3) having to consider not just absolute risk but also the culture about taking and allowing risks in the specific use environment;
(4) considering additional factors such as cybersecurity and privacy.

The goal is to optimise the benefit-risk balance for the end-user(s) by applying the best processes for 'TEVV' (test, evaluation, verification and validation).

The stages of developing and implementing AI MDSW have been described by the National Institute of Standards and Technology (NIST) in the USA as: Data and Input; AI Model; Task and Output; and Application context[38–42]. These have been adapted into 8 phases as represented by the coloured boxes in Fig. 1. Ideally, trustworthy AI should be valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful biases managed[43].

AI tools need to be trained and tested on representative datasets reflecting the context of intended use[44,45], and certain supervised ML algorithms need good-quality labels determined by human annotators for use as reference (or ground truth). The collection of datasets that contain personal information is subject to the requirements of the EU General Data Protection Regulation[46,47]. The use of datasets for developing AI tools and their need for curation necessitate extra attention concerning their validity, intrinsic bias (i.e. by ethnicity, sex, age group, etc.), representativeness of patient populations, geographic distribution of data sources, and quality of the labels.

A comprehensive list of factors that can influence risk is given in Supplementary Information Section 1. For a given AI tool, some or all of these factors will be relevant but usually with variable impact on the overall benefit-risk balance, depending on the application domain and the context

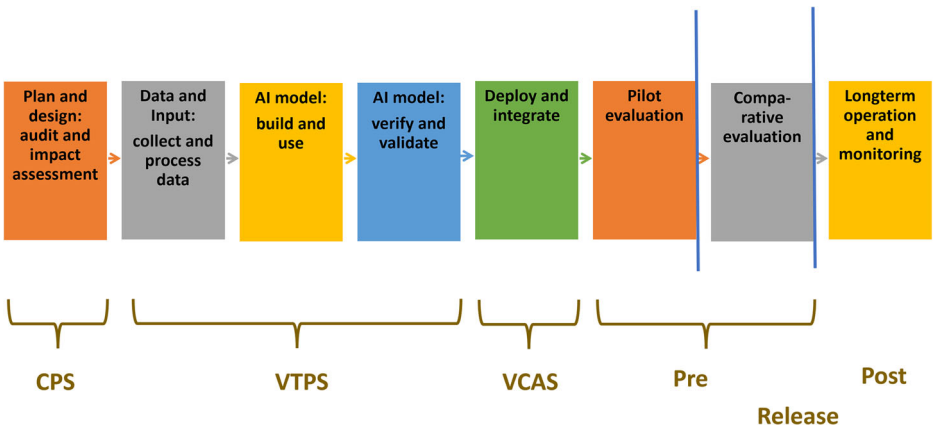


Fig. 1 | Relationship between components of the CORE–MD Risk Score and the stages of development and implementation of AI MDSW. The relationship between components of the CORE–MD Risk Score [CPS Clinical performance score; VTPS Valid Technical performance score; and VCAS Valid clinical performance score] and the stages of development and implementation of AI MDSW (adapted from NIST[42,62,63]) which they have been designed to reflect. The blue vertical lines show the possible timing of CE certification in Europe; depending on the Risk Score, CE certification can be obtained with the pilot (lower risk, certificate with conditions; first line) or with full comparative (higher risk) clinical evaluation (second time-line), always with appropriate post-release evaluation.

of use. A manufacturer should position its AI tool with respect to all these factors, to specify what evidence will be required before approval and what should be collected after release. This information will also inform the end-user when deciding whether to use the tool in the clinical environment.

Balancing safety with access

Individuals and patients should not be exposed to MDSW with unacceptable risks. There have been cases where AI or MDSW applications in healthcare heightened risks for individuals. For example, a US class-action lawsuit against Healthcare United alleged that their AI algorithm ("nH Predict") was issuing wrongful denials of claims for extended care for elderly patients (https://www.forbes.com/sites/douglaslaney/2023/11/16/ai-ethics-essentials-lawsuit-over-ai-denial-of-healthcare/). In another case, involving IBM Watson for Oncology, possible unacceptable risks for medical device software were caused by AI-based systems providing wrong recommendations to doctors for treating cancer[15]. Many AI tools, however, could offer a relevant benefit for unmet needs without conveying significant risk. Not offering such MDSW tools, due to excessive regulatory demands, could disadvantage individuals who would have benefited[48]. Manufacturers could decide not to market their AI medical device in the EU if it is perceived that the burden for CE-marking is too high.

AI tools with a favourable clinical benefit-risk ratio and a low level of risk to individuals and society, as indicated by the scoring system proposed in this advice, could be approved with appropriate graded or stratified evidence. Pre-market studies focused mainly on statistical significance and less on clinical relevance, which could be balanced by more emphasis on gathering data in the post-market phase to support the clinical usefulness. Some low-risk tools, or new releases of existing tools providing usability or other improvements, could be approved without additional studies in patients and with only technical and scientific proof of the desired change or outcome.

A manufacturer might release a lower-risk AI tool with evidence from a study powered for the general target population but not for subpopulations/minority groups. Where the evidence for such groups is limited, patients and users should be made aware of these limitations as appropriate. In contrast, tools with a high risk to individuals or society should undergo extensive clinical evaluation before release, including clinical investigations or trials. If it is difficult or impossible to show benefit, either directly to patients' outcomes or indirectly by improving efficiency for the interactions between HCPs and patients, then any risk, however small, would be unacceptable.

Surveillance of risk

In either case – whether an AI tool is initially approved with graded or more extensive evidence – continued evaluation of its benefit-risk ratio is necessary because of potential drifts over time in its intended or actual use, in the target population, and/or in its verifiability by humans. The representativeness and quality of additional real-world data about the accuracy and performance of the tool need to be demonstrated, evaluated and validated. In contrast to hardware devices, it would be an advantage if such evaluation could be built into the MDSW[49] although that would require specification of new criteria to be assessed by notified bodies. If post-release evidence (algorithmic vigilance)[50] reveals that AI MDSW is negatively influencing the benefit-risk ratio, then stricter regulatory follow-up should ensure that it is withdrawn. Reimbursement decisions would also require reconsideration of new evidence during the post-release phase.

End-users of an AI tool (citizens, patients, and HCPs) should be informed about benefit-risk evaluations and their consequences for certification and access to the market. Transparency is essential for continued trust in a specific tool and in the process as a whole. The quality of submitted evidence should always be high for both pre-release and post-market requirements, and whether data are acquired retrospectively or prospectively, should be determined by analysis of possible benefits and risks.

Estimating risk – scoring system

We propose a simple point-scoring system to estimate the overall risk of an AI tool. It is composed of three parts, that allow assessment of the tool during its whole life cycle. Its components – valid clinical association score (VCAS), valid technical performance score (VTPS), and clinical performance score (CPS) – have been developed using the terms and definitions for categories of evidence given in MDCG guidance 2020-17, which are described in Table 1. Although they are specific to the EU regulatory system,

Table 1 | Components of the CORE–MD AI Risk Score

Criterion and explanation	Level	Score
Valid clinical association score		**VCAS**
How can transparency and oversight be achieved? The MDSW output should have a clear and valid association with its targeted indication (clinical condition or physiological state).	Strong association with easy human oversight and full transparency	1
	Moderate association with difficult human oversight and incomplete transparency	2
	Weak association without the possibility for human oversight and absent transparency	3
Valid technical performance score		**VTPS**
How has the MDSW been validated and tested? The MDSW should be capable of generating technical or analytical output that accurately and reliably reflects the input.	Strong with broad external validation	1
	Moderate with narrow external validation	2
	Weak with only internal validation	3
Clinical performance score		**CPS**
What is the context of the use of the MDSW? Type of disease, condition, disability, or healthcare situation; risk (impact) for the patient.	Non-serious	1
	Serious	2
	Critical	3
What is the medical function of the output? The MDSW should generate clinically relevant output or benefits when it is used as intended.	Inform	1
	Drive	2
	Diagnose or treat	3
	Maximum from the two subscores	6

MDSW medical device software.

The definitions of 'valid clinical association' and 'valid technical performance' were adapted with minor modifications from the IMDRF guidance on Software as a Medical Device (SaMD): Clinical Evaluation (2017). [IMDRF/SaMD WG/N41FINAL:2017][57], and from MDCG 2020–1: Guidance on Clinical Evaluation (MDR) [..] of Medical Device Software[7].

Definitions of the criteria in the clinical performance score were derived from the IMDRF guidance on Software as a Medical Device (SaMD): Clinical Evaluation (2017). [IMDRF/SaMD WG/N41FINAL:2017].

they are well aligned with the IMDRF guidance. We propose that the total score (with possible values from 4 to 12) should be linked to requirements for clinical evaluation before and after an AI tool is approved. Lower values are associated with less risk. The relationships of the three risk scores with the life cycle phases, the timing of the pre- and post-release phases, and the possible timings of CE certification (vertical blue lines) are indicated in Fig. 1.

This approach could seem like an oversimplification but it is intended firstly to help manufacturers, notified bodies, and clinicians to prioritise efforts for evaluating new AI MDSW, and secondly to avoid unnecessarily limiting access to potentially helpful AI tools that are low-risk. The score does not deflect from the need for regulatory appraisal of the entire AI benefit-risk evaluation during the certification process. It does not consider if the manufacturer of an AI tool intends to have its output validated by a 'human-in-the-loop', since that would depend too much on the user's (unknown) expertise and experience. In our view, planning output verification by a human, irrespective of the capability of the human to effectively perform such oversight, is not currently sufficient to designate a medical AI tool as low-risk; there are not yet rigorous and repeatable methods to ensure that explainability is delivered to users (either the HCP or patient) in a manner that truly assists them in recognising poor advice from AI systems, or for the prevention of automaton bias. There are also not yet implementable approaches for evaluating the safety of the explainability of AI MDSW. Should such approaches later be developed, the scoring system can be adapted to take appropriate account of these developments. In all circumstances it should be completely transparent to the end-users, both clinicians and patients, if the AI tool is explainable or not, and how rigorously the validation and testing were performed, in order to support the trust they can put into the results of the tool.

Valid clinical association score (VCAS). The valid clinical association 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"[51]. The clinical association may be characterised by the type of AI model (e.g. supervised or unsupervised), the availability of ground truth to train and test the algorithm, its transparency and explainability, and the possibility for human oversight (see item 4b in the Supplementary Information Section 1).

As an example, in the case of an unsupervised AI model in which data are clustered to find a possible relationship among the extracted features, but without the presence of ground truth, a subscore in this category of 3 (impossible) would be assigned. If following the preliminary association, another and more specific scientific study has been conducted to prove such apparent relationships, then the strength of its results could modulate the relevant VCAS score to being 1 (easy) or 2 (difficult).

An example of effective oversight for a deep learning algorithm that is a 'black box' would be when the output of a diagnostic imaging segmentation tool is verified by a clinician seeing the contour made by the tool, overlaid onto the image that it has analysed; that would merit a subscore in this category of 1 (easy) despite the algorithm itself being uninterpretable.

Valid technical performance score (VTPS). Technical performance is defined as the "Capability of an MDSW to accurately and reliably generate the intended technical/analytical output from the input data". Verification of technical performance is thus by demonstrating that the AI tool accurately, reliably and precisely generates the intended output from the input data, when it is used in the real world in its intended computing and use environments. Technical performance can be documented by standard measures for assessing AI tools, such as accuracy, specificity, sensitivity, area under the receiver-operating characteristic curve, and F1 score, in the presence of a ground truth. Caution is needed with imbalanced datasets, where these standard metrics are overly optimistic and can miss poor performance in low-prevalence conditions[52]. In such cases, measures such as the area under the precision-recall curve provide greater robustness to class imbalance and should be considered instead[53]. Input characteristics are listed in item 4a in Supplementary Information Section 1, and features related to output are given in items 4c–f.

In addition to any metrics, we propose that the grades in the VTPS should reflect the degree of independence between the training data and the data used for testing an AI tool, and the breadth of external testing performed. (see Table 1). Machine learning methods are susceptible to identifying spurious relationships that exist in the training data but are not present in real-world settings[54], resulting in reduced model performance on new data from different settings. Good model performance can be assured only through testing on data acquired from a range of real-world settings[55], which are truly representative of the settings of intended use, and this is reflected by the VTPS (Table 1).

'Internal Validation' means that the performance of the tool has been tested only on data acquired with the same settings (same institution, using the same equipment, interpreted by the same observer as the training group, in the same group of patients, perhaps with bootstrapping) as the training data. This would produce a VTPS of 3. 'Narrow External Validation' implies that the training and testing data were partially differentiated for some of these factors (VTPS = 2), while 'Broad External Validation' signifies that the performance of the AI tool was evaluated using separate training and (re) testing datasets (i.e. acquired using different equipment, from different centres, at different times, interpreted by different observers, in different patient groups, etc). This would generate a VTPS of 1. Thus the VTPS also reflects the risk of bias in the performance of an AI model.

Notwithstanding efforts to validate AI tools before approval, unintended generalisation, shortcut learning[54], biases in the function of the algorithm[56], and other errors in performance often become apparent only when the MDSW is used in the real world for its intended indication but in an unselected population. Drift in the intended purpose and population may occur more easily with MDSW than with other devices, so it is imperative that proper post-release surveillance is conducted. It should document the context of use, the indication for use, and relevant outcomes, at predefined time-points after release. Feedback to regulatory authorities and Notified Bodies should verify the continued performance of the MDSW, and when it becomes necessary to address problems, then interventions should limit use, lead to a recall, or in the EU suspend the certificate of conformity. Such post-release surveillance could also include notifications to end-users to alert them in individual cases if they are using the MDSW outside the validated indication and context.

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"[7,51].

A clinical benefit is always required for unrestricted release to the market; the timing for establishing such benefit, however, depends on the risk involved in exposing patients to the device. The CPS is used to determine such timing, either before or after release.

In 2017, IMDRF guidance recommended that the need for an independent review of clinical investigations and evidence of the benefit of MDSW, before regulatory approval, should be determined firstly according to the function of the software (ranging from informing for a non-serious condition to treating or diagnosing in a critical condition) and secondly to its significance[57]. These features were summarised in two scales ("Definition Statement" and "Impact") but they conflate three characteristics, which are the function of the MDSW (informs/drives/treats), the stage of the clinical condition (non-serious/serious/critical), and its potential impact (none/low/medium/high/catastrophic). Also, the classification system proposed by the IMDRF[58,59] is very context-dependent. The same disease or condition may be acute or chronic, with various levels of severity, and influenced by

comorbidities. A diagnostic tool could be critical when its result determines treatment for a life-threatening disease, while the same tool would be non-serious if used for a chronic non-life-threatening illness or with extensive human oversight. In practice, the application of an AI tool may drift from its original intended purpose ("off-label" use), so it is best from the outset to consider its most critical possible use when determining the risk score.

The CPS consists of two criteria that should be scored separately and then combined (see Table 1); they assess the criticality of the healthcare situation for which the AI tool is intended, and the expected impact of its output. Together they reflect if the AI tool, when used for its recommended indication, achieves the clinical benefit that was claimed as its purpose. The CPS relates to items listed in paragraphs 1–3 and 5–6 in Supplementary Information Section 1. The relevance of human factors is underscored by the emphasis on human oversight (Supplementary Information Section 1, 4b), explainability (Supplementary Information Section 1, 4d), and the evaluation of proper integration in the clinical workflow (Supplementary Information Section 1, 4e).

Discussion

The overall score, which is the sum of the CPS, VTPS and VCAS (Fig. 2), indicates when an extended evaluation and a higher level of clinical evidence would be appropriate before approval, or when a less rigorous assessment

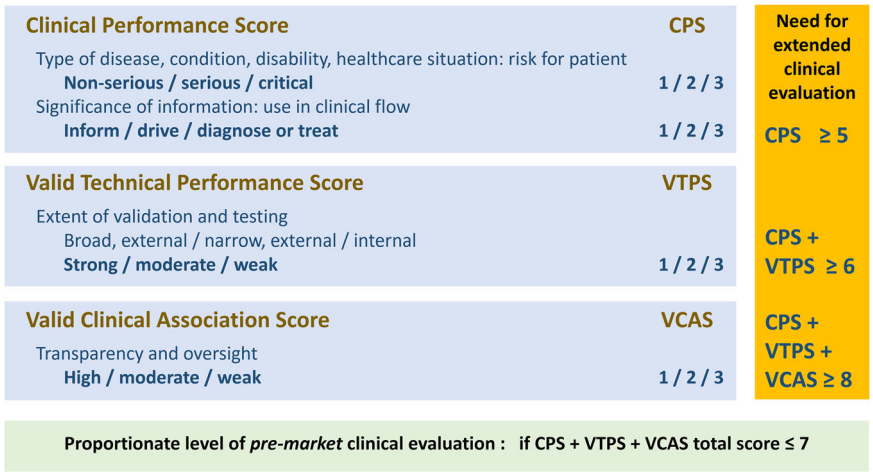
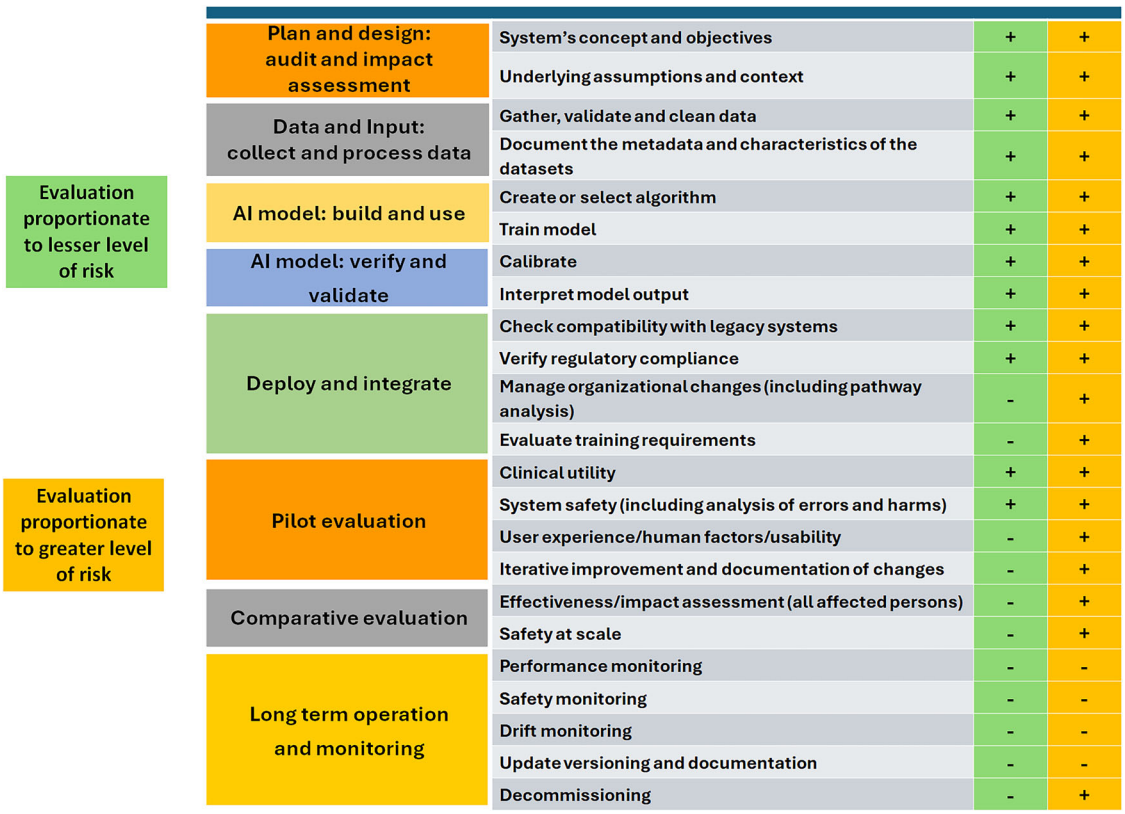
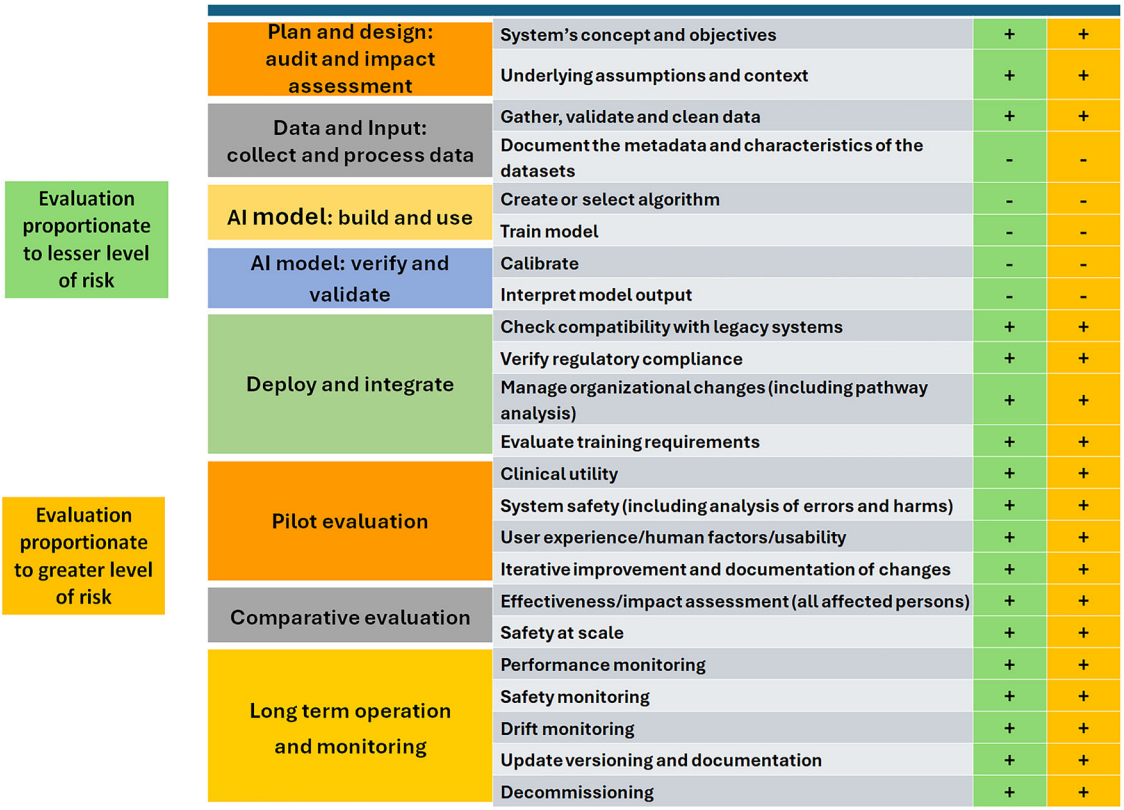


Fig. 2 | Relationships between subtotals and a total of the CORE–MD AI Risk Score and the extent of clinical evaluation recommended before regulatory approval. The CPS score should be estimated first since high scores in both of its parts would mandate more extensive clinical evaluation. Any subtotals or total scores that have values as indicated in the orange box, indicate AI medical devices that merit extensive clinical evaluation before approval. Values falling within the range indicated in the green box will apply to AI medical devices that could be approved for market access after less extensive, appropriate evaluation.



Plan and design: audit and impact assessment	System's concept and objectives	+	+
	Underlying assumptions and context	+	+
Data and Input: collect and process data	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	-	+

Fig. 3 | Recommended items to be evaluated and documented for AI medical device software, before regulatory review, approval, and release – according to the level of the CORE–MD AI Risk Score. Sections 1–8, and the corresponding items, are developed from NIST recommendations[42,62,63]. The green and orange columns correspond respectively to AI medical devices with lower values of the risk score, and those with higher values (see Fig. 2). + indicates an item that should be evaluated *during the pre-market stage*; – indicates an item that does not need to be evaluated at this stage.



Plan and design: audit and impact assessment	System's concept and objectives	+	+
	Underlying assumptions and context	+	+
Data and Input: collect and process data	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	+	+

Fig. 4 | Recommended items to be evaluated and documented for AI medical device software, after regulatory approval and release and according to the level of the CORE–MD AI Risk Score. Sections 1–8, and the corresponding items, are developed from NIST recommendations[42,62,63]. The green and orange columns correspond respectively to AI medical devices with lower values of the risk score, and those with higher values (see Fig. 2). + indicates an item that should be evaluated *during the post-market stage*; – indicates an item that does not need to be evaluated at this stage.

would be proportionate before market release (but perhaps with conditions for collecting more evidence during post-market clinical follow-up). The 10 principles of Good Machine Learning Practices[60] pertain to both circumstances.

A minimum score of 4 would be awarded to an AI tool that has been trained and retested using independent datasets from different populations, gives diagnostic information about a non-serious disease, is fully transparent, completely interpretable and explainable, and is capable of comprehensive human oversight. A low-risk tool with such features can be safely approved with a basic level of pre-market clinical evidence. On the other hand, an autonomous AI system which recommends treatment for a critical condition based on the output of a deep learning algorithm that has not been validated in an independent population and that allows no possibility for human oversight (earning a maximum score of 12), would clearly be very high risk. An AI device of that type cannot be approved safely for market access until it has undergone thorough clinical evaluation, probably including a randomised trial.

Most AI medical devices will fall between these extreme examples. To avoid inappropriately early release of an AI tool with a low overall score but a high score on one of the subsets, extended pre-market evaluation is advised if the CPS score is 5 or more, or if the sum of the CPS and VTPS is 6 or more (see Fig. 2). Thus an AI tool used in a critical situation could fall into the lower-risk category if its function is only to inform (subtotal for CPS = 4); any other function would make it higher risk (≥5). Similarly, an AI decision-support system that suggests a diagnosis or treatment would be lower risk only if its use is restricted to non-serious conditions (CPS = 4). If the technical validation of an AI algorithm has been weak (VTPS = 3) then initial regulatory approval with less extensive pre-market clinical evidence could be considered only if its function claimed by the manufacturer is to inform in a non-serious condition (CPS + VTPS = 5).

Beyond this focus on the CPS and clinical implementation of an AI medical device, the cumulative risk score described in Fig. 2 can also indicate requirements for evaluation across all life cycle stages. These are listed in Figs. 3, 4, opposite eight categories adapted from NIST[30]. The last two columns show items that should be evaluated for lower- and higher-risk AI MDSW, respectively, either *before* market access (Fig. 3) or *after* market access (Fig. 4).

Many official bodies, including the Chinese Authority for Medical Device Evaluation (the National Medical Products Administration, NMPA https://www.cmde.org.cn/xwdt/shpgzgg/gztg/20231107153309174.html), are now engaged in preparing guidelines and recommendations for the use of AI and ML in medical applications; the main sources of documents relevant to the EU system are illustrated in Fig. 5. Within this framework, the CORE–MD project has no official status, but it was funded by the EU with the remit to advise on the clinical evaluation of high-risk medical devices. Representatives from CORE–MD have presented the recommendations in this paper to both the CIE and NT working groups (see Supplementary Information Section 2), with a view to advising on the development of MDCG guidance on the clinical evaluation of AI/ML-enabled medical devices. The scoring system proposed by this paper will be discussed as part of a dedicated work package of CIE aimed at integrating CORE–MD outputs into the European regulatory system. The recommendations have also been presented to one of the chairs of the IMDRF AI/ML working group. A more detailed account of regulatory initiatives relating to AI and ML-enabled medical devices is given in an earlier report from the CORE–MD project[6].

Guidance will need to be developed concerning the methodologies of studies required for each phase. Studies of AI MDSW early in its life cycle aim at showing the stability of the product, while demonstrating safety may be difficult if cohorts are small. Later, comparative studies could make use of

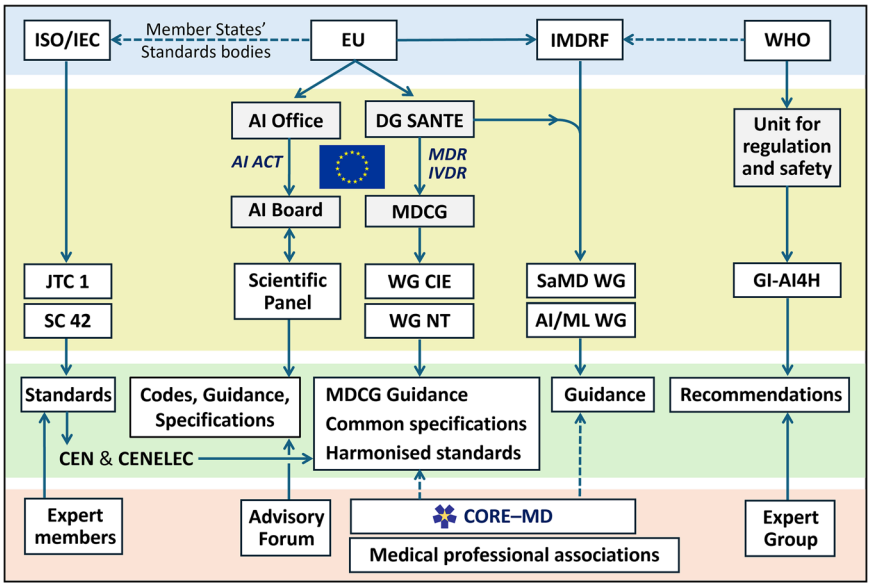


Fig. 5 | Schematic representation of major bodies involved in producing guidance relevant to the EU for approving AI medical device software. In the European Union (EU), the European Commission is responsible for proposing legislation; its Directorate General for Health and Food Safety (DG SANTE) is responsible for the harmonised implementation of the Medical Device Regulation (MDR) and the In Vitro Diagnostic Medical Device Regulation (IVDR) in consultation with member states at the Medical Device Coordination Group (MDCG). The MDR authorises the MDCG to develop device standards, common specifications, and product-specific guidance. Its working groups on Clinical Investigation and Evaluation (CIE) and New Technologies (NT) share interests in the clinical evaluation of medical device software and AI/ML-enabled medical devices. The AI Office established within DG CNECT (Communications Networks, Content and Technology) will manage the implementation of the AI Act, with member states at the AI Board and with advice from both a Scientific Panel and an Advisory Forum. The AI Act provides authority for the production of codes of practice, guidance documents, and specifications, within particular fields. Member states of the EU – independently from their device regulators – are members of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), through their national standards bodies. The ISO and IEC have a joint technical committee (JTC 1) on Information Technology; its sub-committee (SC) 42 has the responsibility to prepare guidance on Artificial Intelligence. ISO/IEC standards may be harmonised with EU legislation by their European counterpart organisations CEN (the European Committee for Standardization) and CENELEC (the European Electrotechnical Committee for Standardization), on request from the European Commission. The EU is a member of the International Medical Device Regulators Forum (IMDRF) where it collaborates with other regulatory jurisdictions to prepare joint recommendations. The IMDRF has established working groups for Software as a Medical Device (SaMD) and for AI and ML. Finally, within the World Health Organization (WHO), which has official observer status at the IMDRF, there is a unit concerned with the regulation and safety of medical devices; it collaborates with the International Telecommunication Union (ITU) in a Global Initiative on AI for Health (GI-AI4H) that has also produced guidance on regulatory standards.

real-world approaches, large simple trials, and adaptive designs[61]. Retrospective data can be used for the initial training, testing, and validation of an AI tool, but prospective clinical investigations will always be required in the appropriate phases. Medical device Expert Panels could play a crucial role in establishing guidance for clinical evaluation, with participation from patients and citizens especially when rules for the clinical evaluation of MDSW in lower-risk classes are being considered; these CORE–MD recommendations are relevant especially to high-risk AI MDSW.

Evaluation of the potential clinical utility of the CORE–MD AI Risk Score will need to assess if it is equally applicable to different AI/ML devices used in different contexts and populations. The use of a new tool may drift, and it may be used 'off-label'. All AI/ML devices need to be evaluated to determine their generalisability; a similar consideration may be important for this score.

Conclusions

Using a new and simple scoring system for the assessment of risk, we propose a benefit-risk approach to guide requirements for the clinical evaluation of AI Medical Device Software, taking into account the entire life cycle of the MDSW and addressing regulatory requirements and the clinical evidence needed for trusted use of these devices by patients and caregivers.

By combining regulatory and clinical requirements into one workflow, and by focusing on the need for real-world evidence for AI MDSW, including an analysis of the risks involved in human-machine interactions, we offer a more streamlined approach which can lead to a proportionate implementation of the MDR requirements, alleviating some of the concerns about limiting innovation. By emphasising the post-release phase, any changes or drift in AI MDSW in the clinical environment can be addressed when and where they occur.

The approach that we recommend should now be evaluated scientifically, and if its utility is confirmed then it could serve as a valuable contributary resource for future EU regulatory guidance. The recommendations have been developed robustly, in consultation with regulatory authorities, and further discussions are planned.

Received: 12 July 2024; Accepted: 15 January 2025;
Published online: 06 February 2025

References

1. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
2. Hunter, D. J. & Holmes, C. Where medical statistics meets artificial intelligence. *N. Engl. J. Med.* **389**, 1211–1219 (2023).
3. World Health Organization. Regulatory considerations on artificial intelligence for health. https://www.who.int/publications/i/item/9789240078871 (2023).
4. Bitkina, O. V., Park, J. & Kim, H. K. Application of artificial intelligence in medical technologies: a systematic review of main trends. *Digit. Health* **9**, 20552076231189331 (2023).
5. Ellahham, S., Ellahham, N. & Simsekler, M. C. E. Application of artificial intelligence in the health care safety context: opportunities and challenges. *Am. J. Med. Qual.* **35**, 341–348 (2020).
6. Fraser, A. G. et al. Artificial intelligence in medical device software and high-risk medical devices – a review of definitions, expert

recommendations and regulatory initiatives. *Expert Rev. Med. Devices* **20**, 467–491 (2023).
7. European Commission. Medical Devices Coordination Group Document MDCG 2020-1. Guidance on Clinical Evaluation (MDR) / Performance Evaluation (IVDR) of Medical Device Software. https://health.ec.europa.eu/system/files/2020-09/md_mdcg_2020_1_guidance_clinic_eva_md_software_en_0.pdf (2020).
8. Fraser, A. G. et al. Improved clinical investigation and evaluation of high-risk medical devices: the rationale and objectives of CORE-MD (Coordinating Research and Evidence for Medical Devices). *Eur. Heart J. Qual. Care Clin. Outcomes* **8**, 249–258 (2022).
9. Balahur-Dobrescu, A. et al. Data quality requirements for inclusive, non-biased and trustworthy AI. https://doi.org/10.2760/365479, JRC131097 (Publications Office of the European Union, Luxembourg, 2022).
10. Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B. & Richardson, W. S. Evidence based medicine: what it is and what it isn't. *BMJ* **312**, 71–72 (1996).
11. Rozich, J. D. et al. Standardization as a mechanism to improve safety in health care. *Jt. Comm. J. Qual. Saf.* **30**, 5–14 (2004).
12. Kahneman, D., Sibony, O. & Sunstein, C. R. *Noise: a Flaw in Human Judgment* (Little, Brown Spark, New York, 2021).
13. McNeese, N. J., Demir, M., Cooke, N. J. & Myers, C. Teaming with a synthetic teammate: insights into human-autonomy teaming. *Hum. Factors* **60**, 262–273 (2018).
14. Chesterman, S. Through a glass, darkly: artificial intelligence and the problem of opacity. *Am. J. Comp. Law* **69**, 271–294 (2021).
15. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at http://arxiv.org/abs/1702.08608 (2017).
16. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
17. Nicholson Price, W. Big data and black-box medical algorithms. *Sci. Transl. Med.* **10**, eaao5333 (2018).
18. Tschandl, P. et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatol.* **155**, 58–65 (2019).
19. Tschandl, P. Risk of bias and error from data sets used for dermatologic artificial intelligence. *JAMA Dermatol.* **157**, 1271–1273 (2021).
20. Gaube, S. et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* **4**, 31 (2021).
21. Tschandl, P. et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
22. Agarwal, N., Moehring, A., Rajpurkar, P. & Salz, T. *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology*. Working paper 31422 https://doi.org/10.3386/w31422 (National Bureau of Economic Research, 2023).
23. Yu, F. et al. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat. Med.* **30**, 837–849 (2024).
24. Vasey, B. et al. Association of clinician diagnostic performance with machine learning-based decision support systems: A systematic review. *JAMA Netw. Open* **4**, e211276 (2021).
25. Shneiderman, B. Human-centered artificial intelligence: reliable, safe & trustworthy. *Int. J. Hum. Comput. Interact.* **36**, 495–504 (2020).
26. Dey, D. et al. Proceedings of the NHLBI workshop on artificial intelligence in cardiovascular imaging: translation patient care. *JACC Cardiovasc. Imaging* **16**, 1209–1223 (2023).
27. Gilbert, S. et al. Algorithm change protocols in the regulation of adaptive machine learning-based medical devices. *J. Med. Internet Res.* **23**, e30545 (2021).
28. Coiera, E. The last mile: where artificial intelligence meets reality. *J. Med. Internet Res.* **21**, e16323 (2019).
29. Lyell, D., Coier, A. 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.* **28**, e100301 (2021).
30. Health Sciences Authority (Singapore). Regulatory guidelines for software medical devices - A lifecycle approach. Revision 2.0 https://www.hsa.gov.sg/docs/default-source/hprg-mdb/guidance-documents-for-medical-devices/regulatory-guidelines-for-software-medical-devices---a-life-cycle-approach_r2-(2022-apr)-pub.pdf (2022).
31. Zhang, J. et al. Moving towards vertically integrated artificial intelligence development. *NPJ Digit. Med.* **5**, 143 (2022).
32. Gerke, S., Babic, B., Evgeniou, T. & Cohen, I. G. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit. Med.* **3**, 53 (2020).
33. Daly, A. et al. *Artificial Intelligence, Governance and Ethics: Global Perspectives*. Paper No. 2019/033. https://doi.org/10.2139/ssrn.3414805 (Chinese University of Hong Kong Faculty of Law Research, 2019).
34. Biasin, E., Brešić, D., Kamenjašević, E. & Notermans, P. Analysis of ethics, privacy, and confidentiality constraints. SAFECARE project deliverable 3.9. Lead Author: KUL. https://www.safecare-project.eu/wp-content/uploads/2020/02/Analysis-of-Ethics-Privacy-and-Confidentiality-Restraints.pdf (2018).
35. Faden, R. R. et al. An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. *Hastings Cent. Rep.* **43**, Spec No:S16–27 (2013).
36. High-Level Expert Group on Artificial Intelligence set up by the European Commission. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (2019).
37. Ayers, J. W., Desai, N. & Smith, D. M. Regulate artificial intelligence in health care by prioritizing patient outcomes. *JAMA* **331**, 639–640 (2024).
38. NIST. *AI Risk Management Framework Playbook-MAP* https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook (2023).
39. NIST. *AI Risk Management Framework Playbook-MEASURE* https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook (2023).
40. NIST. *AI Risk Management Framework Playbook-MANAGE* https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook (2023).
41. NIST. *AI Risk Management Framework Playbook-GOVERN* https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook (2023).
42. NIST. *Risk Management Framework Quick Start Guide. Roles and Responsibilities Crosswalk* https://nist.gov/rmf (2021).
43. NIST. *Privacy Framework Core* https://www.nist.gov/privacy-framework (2020).
44. Bernal-Delgado, E. et al. Report on architecture and infrastructure options to support EHDS services for secondary use of data. https://tehdas.eu/app/uploads/2023/03/tehdas-report-on-architecture-and-infrastructure-options-to-support-ehds-services.pdf (2023).
45. Newlands, G., Lutz, C. & Fieseler, C. Trading on the unknown: scenarios for the future value of data. *Law Ethics Hum. Rights* **13**, 97–114 (2019).
46. Andanda, P. Towards a paradigm shift in governing data access and related intellectual property rights in big data and health-related research. *IIC Int. Rev. Intellect. Prop. Compet. Law* **50**, 1052–1081 (2019).
47. Starkbaum, J. & Felt, U. Negotiating the reuse of health-data: Research, Big Data, and the European General Data Protection Regulation. *Big Data Soc*. https://doi.org/10.1177/2053951719862594 (2019).
48. Kadakia, K. T., Kramer, D. B. & Yeh, R. W. Coverage for emerging technologies — bridging regulatory approval and patient access. *N. Engl. J. Med.* **389**, 2021–2024 (2023).

49. Welzel, C. et al. Holistic human-serving digitization of health care needs integrated automated system-level assessment tools. *J. Med. Internet Res.* **25**, e50158 (2023).
50. Balendran, A., Benchoufi, M., Evgeniou, T. & Ravaud, P. Algorithmovigilance, lessons from pharmacovigilance. *NPJ Digit. Med.* **7**, 270 (2024).
51. European Commission. Guidelines on the qualification and classification of stand alone software used in healthcare within the regulatory framework of medical devices. Medical Devices Guidance Document MEDDEV 2.1/6 (2016)European Commission. Guidelines on the qualification and classification of stand alone software used in healthcare within the regulatory framework of medical devices. Medical Devices Guidance Document MEDDEV 2.1/6 (2016).
52. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **6**, 27 (2019).
53. Mosquera, C., Ferrer, L., Milone, D. H., Luna, D. & Ferrante, E. Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance. *Eur. Radiol*. https://doi.org/10.1007/s00330-024-10834-0 (2024).
54. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
55. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
56. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
57. International Medical Device Regulators Forum. Software as a Medical Device Working Group. Software as a Medical Device (SaMD): Clinical Evaluation. IMDRF/SaMD WG/N41FINAL:2017. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-170921-samd-n41-clinical-evaluation_1.pdf (2017).
58. International Medical Device Regulators Forum. IMDRF/SaMD WG/N23 FINAL: 2015 - Software as a Medical Device (SaMD): Application of Quality Management System. (2015).
59. International Medical Device Regulators Forum. Machine Learning-enabled medical devices-a subset of artificial intelligence-enabled medical devices: key terms and definitions. IMDRF AIMD Working Group. (2021).
60. US Food and Drug Adminstration, Health Canada, and the Medicines & Healthcare Products Regulatory Authority. Good machine learning practice for medical device development: guiding principles. https://www.fda.gov/media/153486/download (2021).
61. Fleetcroft, C., McCulloch, P. & Campbell, B. IDEAL as a guide to designing clinical device studies consistent with the new European medical device regulation. *BMJ Surg. Interv. Health Technol.* **3**, e000066 (2021).
62. Warraich, H. J., Tazbaz, T. & Califf, R. M. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. *JAMA* https://doi.org/10.1001/jama.2024.21451 (2024).
63. FDA. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)—discussion paper and request for feedback https://www.fda.gov/files/medical%20devices/published/USFDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf (2024).

Acknowledgements

The CORE–MD project received funding from the European Union Horizon 2020 Research and Innovation programme, for a Coordination & Support Action, under grant agreement No. 945260. We thank all the colleagues who took part in the Delphi project and consultations.

Author contributions

F.R. co-designed the AI MDSW work package in the CORE–MD programme; led the online and on-site meetings; drafted and revised the deliverables and publications output of the work package, including this manuscript. E.B. contributed legal expertise to the work package discussions; drafted and revised parts of the text outputs of the work package, including this manuscript. N.B. contributed to the discussions leading to this manuscript; drafted and revised parts of the text outputs of the work package, including this manuscript. E.C. contributed to the concepts and deliverables of the work package; drafted and revised parts of the text outputs of the work package, including this manuscript. R.D. contributed to the concepts and deliverables of the work package; drafted and revised parts of the text outputs of the work package, including this manuscript. S.G. contributed to the concepts and deliverables of the work package; drafted and revised parts of the text outputs of the work package, including this manuscript. E.K. contributed legal expertise to the work package discussions; drafted and revised parts of the text outputs of the work package, including this manuscript. G.M. contributed to the concepts and deliverables of the work package; drafted and revised parts of the text outputs of the work package, including this manuscript. G.O. contributed to the concepts and deliverables of the work package; drafted and revised parts of the text outputs of the work package, including this manuscript. J.B.R. organised the Delphi-like procedure reported in this manuscript; drafted and revised parts of the text outputs of the work package, including this manuscript. B.V. contributed to the concepts and deliverables of the work package; drafted and revised parts of the text outputs of the work package, including this manuscript. A.G.F. led the CORE–MD consortium; obtained the EU funding; co-designed the AI MDSW work package in the CORE–MD programme; drafted and revised the deliverables and publications of the work package, including this manuscript. All authors have read the paper, approved the submission of the paper and agreed to be responsible for their own contribution and subscribe to the ethics guidelines to ensure compliance with the accuracy and integrity of the work.

Competing interests

F.R.: Occasional travel and accommodation costs to CORE–MD meetings were supported within the EU Horizon 2020 grant budget (2022-204); senior associate editor European Heart Journal, Cardiovascular Imaging. E.B.: external collaborative expert at the European Medicine Agency between 2023 and 2024 (unpaid collaboration); Occasional travel and accommodation costs to CORE–MD meetings were supported within the EU Horizon 2020 grant budget (2022-204). N.B.: Honorarium for presentation at Great Wall Cardiac Congress; ESC paid Article Processing Costs for one article in the European Heart Journal – Digital Health; Heart Rhythm Society paid travel costs for presentation at HRX; ESC Vice-chair Digital Health Committee; Editor-in-Chief of European Heart Journal – Digital Health. E.C.: Research contract with Luxottica S.p.A, Payment to my institution; Research contract with Croce Rossa Italiana, Payment to my institution; Project evaluator for the Autonomous Province of Trento, Italy; Honoraria for presentations from Dynamicom Education Srl, Summeet Srl, AIM Italy S.r.l.; Member of the Regulatory Affairs committee of the European Society of Cardiology. R.D.: minority shareholder in myocardium.AI; Consulting fees were paid directly to him. The company develops AI products for cardiac MR image analysis and is in the process of applying for regulatory approval. S.G.: German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) through the European Union-financed NextGenerationEU programme under grant number 16KISA100K, project PATH—'Personal Mastery of Health and Wellness Data'. The European Commission under the Horizon Europe Programme, as part of the projects CYMEDSEC (101094218) and ASSESS-DHT (101137347). S.G. has or has had consulting relationships with Una Health GmbH, Lindus Health Ltd.; Flo Ltd, Thymia Ltd., FORUM Institut für Management GmbH, High-Tech Gründerfonds Management GmbH, Ada Health GmbH, and he holds share options in Ada Health GmbH. Support and honoraria for TK Health Economics Forum at MEDICA. Advisory Group member of the EY-coordinated 'Study on Regulatory Governance and Innovation in the Field of Medical Devices' conducted on behalf of the DG SANTE of the

European Commission. Share options in Ada Health GmbH. S.G. is a News and Views Editor for npj Digital Medicine. S.G. played no role in the internal review or decision to publish this article. G.M.: Occasional travel and accommodation costs to CORE–MD meetings were supported within the EU Horizon 2020 grant budget (2022-204). Health Products Regulatory Authority, Source of employment. G.O.: Health Products Regulatory Authority, Source of employment. J.B.R.: none. B.V.: Berrow Foundation Lord Florey scholarship. A.G.: Chair, Regulatory Affairs Committee, Biomedical Alliance in Europe.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01459-8.

Correspondence and requests for materials should be addressed to Frank E. Rademakers.