MINISTRY OF HEALTH
For a healthier life

Principles for the Evaluation Of AI-Based Interventional Clinical Trials

Ministry of Health, Israel

February 2026 – Version 2.0

Table of Contents:

Preface

The evolution of Artificial Intelligence (AI) within the dynamic landscape of medical technologies is ushering in a new era of possibilities in healthcare. AI-based tools are increasingly integrated into the healthcare system, leading to a higher frequency of requests for clinical trials involving these tools. Consequently, there is a growing need to establish guiding principles for the review and approval of such requests, ensuring safety, efficacy, and adherence to ethical standards.

The purpose of this document is to outline the guiding principles for evaluating applications for interventional clinical trials involving AI-based tools[1], tailored to the unique considerations inherent in these technologies. These principles are intended to provide comprehensive recommendations and guidelines for Institutional Review Board (IRB), also referred to in Israel as the IRB, whose role is to assess submissions for trials involving AI-based products and to enhance decision-making processes by highlighting issues unique to these technologies. This is done while promoting the values of safety, privacy, accountability, fairness, disclosure, transparency, and explainability in the use of AI models.

It should be emphasized that this document does not define when a prospective interventional ("active") study must be conducted; rather, it focuses on the issues that must be examined when evaluating applications for such studies. The table of principles below guides a balanced approach to evaluating AI-based products in prospective interventional studies. By adopting these guidelines, committees can promote high standards of scientific rigor and ethical practice, ultimately advancing the improvement of patient care through these innovative technologies.

It is further clarified that this document does not address autonomous systems, which require the examination of additional aspects not included within this framework.

Given the rapid pace of AI technology development, the evolving understanding of ethical and safety considerations, and the progress in international regulatory infrastructure, we view this as a "living document" subject to updates. We welcome comments and feedback at: samd@moh.gov.il.

[1] Throughout this document, the terms “interventional”, “prospective interventional”, and “active” are used interchangeably to refer to prospective clinical studies in which the AI-based tool may influence clinical decision-making, unless explicitly stated otherwise.

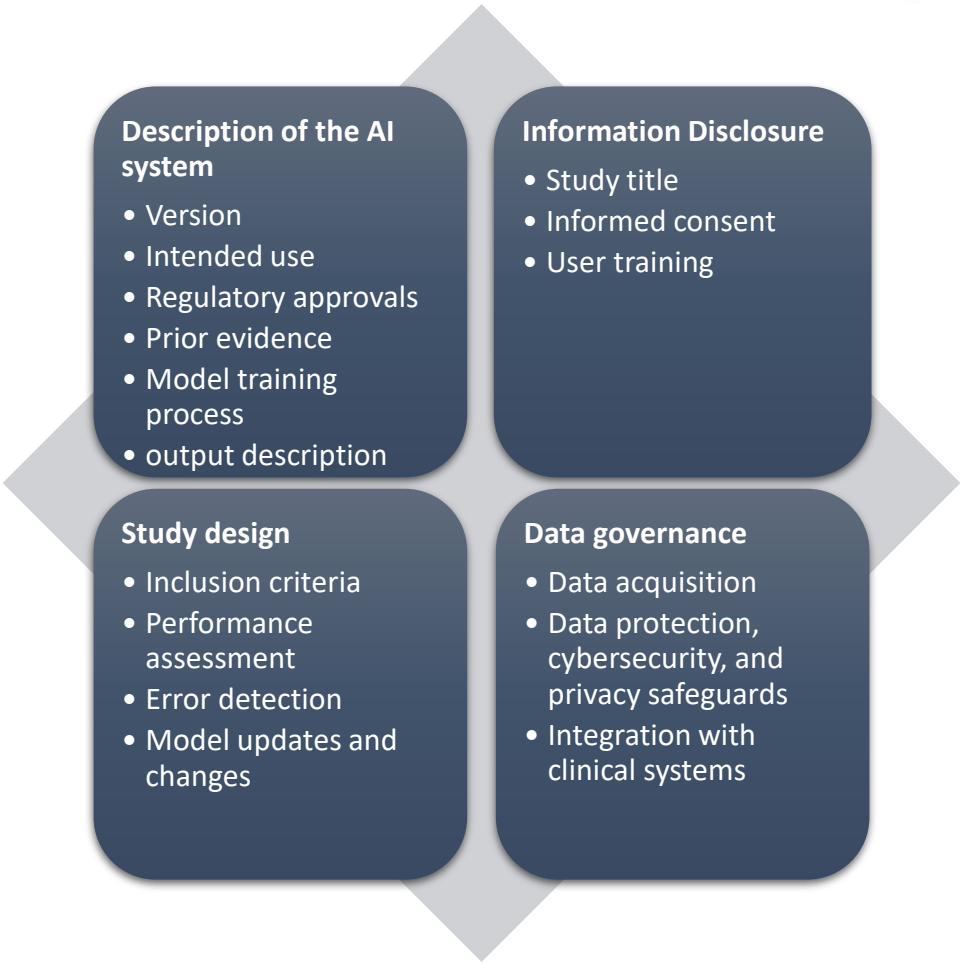
Executive Summary

This document presents guiding principles for the evaluation of applications for interventional clinical trials involving AI-based tools. Its objective is to provide Institutional Review Boards (IRBs) (referred to in Israel as “Helsinki Committees”) with a structured framework for the ethical, scientific, clinical, and safety assessment of studies involving AI-based interventions. This framework encourages the responsible and transparent use of innovative technologies within the healthcare system.

The principles in this document were developed based on international guidelines and a broad consultation process with representatives from healthcare institutions, academia, ethics, and regulation. They are designed to support a risk-based approach tailored to the complexity and novelty of these technologies, while upholding the values of safety, privacy, accountability, and fairness.

The document is intended for IRBs, researchers, and all professional entities involved in the submission, review, and execution of AI-based research.

While the document presents various types of AI-based research (see Types of studies across the AI), it focuses on prospective interventional study where the AI tool may actively influence clinical decision-making. The document includes a table detailing 16 key issues, divided into four categories, arranged according to the logical sequence of evaluating an AI-based intervention throughout the research stages. from understanding the system's characteristics and the product under evaluation, through the management of the data it relies on, to study design and information accessibility for stakeholders:



Following the table, the document includes clarifications regarding the principles and supplementary guidelines for each of the issues, to provide a deeper understanding and practical guidance for the committees (see section: Detailed Guiding Principles for Evaluating Interventional Research Involving AI Tools). Additionally, the document includes a dedicated chapter dealing with relevant terminology. The following are the main questions focused on in the document:

#	Question	Category
1	Does the clinical trial protocol and trial registration explicitly state that the intervention includes an AI-based component, and what is its degree of autonomy?	Information Disclosure
2	When reviewing the Informed Consent Form, does it include information on the use and degree of intervention of the AI system, and explain the purpose of use, mode of operation, risks and benefits, data handling, and reference to local regulation?	Information Disclosure

#	Question	Category
3	What is the version of the model/tool being used?	Description of the AI System
4	Has the Intended Use of the AI tool been detailed?	Description of the AI System
5	What are the inclusion and exclusion criteria at the participant level and for the input data to the model during the study? What is the rationale for these criteria? Do these criteria match the tool's target population?	Study Design
6	Are there prior regulatory approvals for the tool?	Description of the AI System
7	Is there prior information and evidence for intervention using the specific AI tool or another AI tool for the same indication/clinical purpose?	Description of the AI System
8	What is the process for collecting/acquiring the data fed into the system, and how is the data processed and stored?	Data Governance
9	What mechanisms and tools exist for data protection, cyber security, and privacy?	Data Governance
10	How was the model trained?	Description of the AI System
11	How will the model's performance be tested?	Study Design
12	How will errors be identified and handled?	Study Design
13	Does the tool require integration with other systems in the clinical environment?	Data Governance
14	What does the output of the AI intervention look like? How does it contribute to the decision-making process or any other clinical activity?	Description of the AI System
15	Is the training process for using the AI system described in the protocol?	Information Disclosure
16	Is there a reference to reporting significant changes in the model to the committee? Are the criteria for early study termination well-described in the protocol?	Study Design

Background

Unique Characteristics of Artificial Intelligence Technologies

While similarities exist between medical software and AI-based tools, this document focuses on the unique characteristics of the latter, including dependence on data quality, the capacity for continuous learning and performance updates, and the occasional difficulty in explaining outcomes due to algorithmic complexity. Consequently, evaluating these products requires a high level of disclosure and transparency regarding the sources and databases used for training, training techniques, model limitations, performance data, and existing control pathways.
Another key aspect is the increasing prevalence of Clinical Decision Support Systems (CDSS). Although they do not intervene directly in treatment, they are becoming more influential due to their ability to process big data and provide insights that actively influence how clinical information is analyzed, how interventions are prioritized, and how decisions are made by the care team. In some cases, the difficulty in explaining how the software reached a certain conclusion creates significant challenges in validating its outputs, highlighting the need for clinical (experimental) evaluation of the system's usage outcomes.
Furthermore, recent years have seen a breakthrough in Generative AI, and particularly in Large Language Models (LLMs), which are gradually entering the clinical workspace. These tools pose unique questions regarding dependence on external sources, traceability, and explainability. Specific aspects regarding the use of such technologies will be highlighted throughout the document.

Document Formulation Process

This document is the second and updated version of the document published in October 2023. It was formulated by a multi-disciplinary committee established to draft the principles and assist in evaluating Software as a Medical Device (SaMD) arriving at the Clinical Trials Department of the Ministry of Health. This committee includes representatives from medical institutions, academia, and experts in medicine, technology, ethics, and regulation. The committee's work relied on the principles of

the SPIRIT-AI checklist for clinical trials[2] and complementary documents, including the OPTICA[3] document published by Clalit Health Services. The formulation process included ongoing dialogue with IRBs and was opened for broad public consultation to receive and integrate further comments. Subsequently, the principles were also drafted based on real-world case studies used to test the applicability of the guidelines in diverse contexts.

How to Use the Presented Principles

The principles presented below are displayed as questions, emphasizing the importance of evaluating the benefit of AI-based interventions against the risks to patient safety and privacy. The principles highlight the need to examine the quality of information used for training and testing the models, as well as potential biases originating from the training database. Additionally, reference is made to data flow processes from the environment to the models and back, their mode of operation, and the manner of using these technologies, with an emphasis on Human-Machine Interaction and the training required for informed and supervised use while maintaining ethical standards. Most of this information will be provided by the applicants, with the questions aiming to supply the information required for a qualitative review by the committee.

Recognizing that not all AI applications are equal in terms of risk profile or clinical benefit, the examination of each application should be commensurate with the complexity and novelty of the evaluated technology and the clinical context in which it is to be applied. IRBs should exercise discretion in applying these principles,

[2] Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension | Nature Medicine SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence) is an extension of SPIRIT 2013 aimed at improving the transparency and completeness of clinical trial protocols involving AI-based interventions. It was developed by an international group of experts through a consensus process that included a literature review, a Delphi survey, and consensus meetings, and was published in Nature Medicine in 2020. This extension includes 15 additional items that were identified through the consensus process as particularly important for the evaluation of such protocols. These items cover, among other aspects, detailed descriptions of the AI intervention, instructions for use, handling of input and output data, human–AI interactions, and the analysis of error cases, with the goal of promoting transparency and completeness in clinical trial protocols.

[3] https://ai.nejm.org/doi/abs/10.1056/AIcs2300269

consulting with experts in the field, and evaluating applications based on perceived risk and benefit.

In recent years, AI governance committees have been established within healthcare organizations, focusing on risk management and the responsible implementation of these systems. IRBs can seek input from these committees as a supporting professional body when reviewing submissions for clinical trials that involve the use of artificial intelligence. To classify the risk of the tool in the study, one may refer to IMDRF documents dealing with risk management[4]. If the healthcare organization implements a structured process for AI regulation regarding additional stages of the AI product lifecycle (beyond the research stage), information may be collected in an alternative format, provided that format addresses all issues detailed in the table.

It is noted that while this document is intended for the review of interventional clinical trials, many issues raised are also relevant to earlier and later stages of the research, development, and implementation process. The document echoes the contemporary view regarding guiding principles for developing Machine Learning technologies, relevant to all lifecycle stages of ML development[5] in healthcare. Familiarity with and adherence to these principles during the early stages of model development may assist developers in establishing high-quality workflows, maintaining transparency and control, and bringing tools to a higher level of readiness for the interventional research stage.

[4] "Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations (imdrf.org)
[5] https://www.gov.il/en/pages/digital-medical-technology-gmlp-1

Detailed Guiding Principles for Evaluating Interventional Research Involving AI Tools

The table below presents 16 key issues to be considered when evaluating interventional research involving AI tools. The topics are divided into four main categories, representing a logical sequence of evaluating the intervention: from understanding the system itself, through data processing and integration into organizational systems, to how it is tested and reflected to various stakeholders within the study:

- **Description of the AI System:** This category deals with understanding product characteristics: what model is used, which version is tested, what is the Intended Use, what regulatory approvals or prior evidence exist, how the model was trained, and what are its output characteristics. The goal is to establish a complete picture of the nature of the intervention and its functionality.
- **Data Governance:** Focuses on the data infrastructure underpinning the model's operation: how data is collected, from which sources, under what conditions it is stored and processed, and how data protection, cyber security, and privacy principles are maintained. This category emphasizes that data quality and safety are fundamental conditions for the responsible use of AI technologies.
- **Study Design:** Refers to methodological aspects of the study itself: inclusion and exclusion criteria, how model performance is tested during the study, and mechanisms for dealing with potential changes, whether in performance, the model itself, or environmental conditions. This category is designed to help the committee evaluate the adequacy of the study design and the ability to verify adherence to ethical and safety principles while maintaining result reliability.
- **Information Disclosure:** Deals with transparency aspects surrounding AI usage. It includes questions on how the AI product usage is clearly stated in the study title, how information is conveyed to patients, and how explanations and training are provided to teams using the system. The goal is to ensure full transparency for all relevant stakeholders as part of the need to establish accountability and build trust in the technology.

In each of the issues in the table, guiding questions are presented, aiming to stimulate discretion and refine the key examination points in each area. Further explanation regarding the issues appears in the expansion chapter below, containing information that includes contexts, examples, and additional ethical and technical aspects.

#	Question	Category	Considerations
1	Does the clinical trial protocol and trial registration explicitly state that the intervention includes an AI-based component, and what is its degree of autonomy?	**Information Disclosure**	Reference should be made to the degree of autonomy/automation of the tool (fully autonomous, partially autonomous, or non-autonomous[6]).
2	When reviewing the Informed Consent Form, does it include information on the use and degree of intervention of the AI system, and explain the purpose of use, mode of operation, risks and benefits, data handling, and reference to local regulation?	**Information Disclosure**	The informed consent form should be written in clear, understandable language and include a concise explanation of the implications of using AI in the studied context.
3	What is the version of the model/tool being used?	**Description of the AI System**	
4	Has the Intended Use of the AI tool been detailed?	**Description of the AI System**	The intended use should address: • Medical purpose

[6] IMDRFSaMD WGN81 DRAFT 2024, Medical Device Software Considerations for Device and Risk Characterization - final draft.pdf

#	Question	Category	Considerations
			• Target patient population • Intended users (clinicians, patients, general public, etc.) • Use environment • Out-of-scope use cases • Description of inputs and outputs • Role of outputs in the medical process • Mode of use and integration into clinical workflow and degree of autonomy • Learning/update characteristics (planned or continuous)
5	What are the inclusion and exclusion criteria at the participant level and for the input data to the model during the study? What is the rationale for these criteria? Do these criteria match the tool's target population?	**Study Design**	• Do the inclusion criteria reflect the clinical indication (Question 4) and the characteristics of the data used to train the model (Question 10)? • Was fairness and adequate representation explicitly assessed when defining the inclusion and exclusion criteria? • Do the criteria include conditions related to the availability, quality, and completeness of the input data required to operate the tool? • Where applicable, was the maintenance of the inclusion criteria examined when the model is retrained during the study (see Question 15)?
6	Are there prior regulatory approvals for the tool?	**Description of the AI System**	Where prior regulatory authorization exists: • Do the approvals explicitly cover the AI component of the product? • Have any changes been made to the product or to the model since the authorization was granted?

#	Question	Category	Considerations
			• What clinical indication is specified in the authorization?
7	Is there prior information and evidence for intervention using the specific AI tool or another AI tool for the same indication/clinical purpose?	**Description of the AI System**	• Is there prior evidence regarding the AI tool for the clinical objective of the study? • What type of evidence has been presented (retrospective, prospective, or other)? • Under what conditions and in what setting were the data supporting this evidence generated?
8	What is the process for collecting/acquiring the data fed into the system, and how is the data processed and stored?	**Data Governance**	• Is the data collection process comparable to the data used for model training and to the setting in which the intervention is intended to be deployed? • Is there human–AI interaction in the handling or processing of input data? If so, is a specific level of expertise required from participants or users? • Is there an assessment and management plan for missing data, outliers, low-quality input data, inaccessible data, or data that were entered incorrectly? • Whether and how will the stability of the input data be monitored over time? • Is there a risk management plan in place to prevent or mitigate bias in the input data?

#	Question	Category	Considerations
9	What mechanisms and tools exist for data protection, cyber security, and privacy?	**Data Governance**	• Has the organization's designated officer for data protection, cybersecurity, and privacy reviewed compliance with the relevant organizational and regulatory requirements? • In cases where a third-party tool is used (e.g., an LLM), is the approach for safeguarding participant privacy and limiting the use of data clearly described?
10	How was the model trained?	**Description of the AI System**	Where known, information should be provided on the following aspects: • Which model is used, who is the manufacturer, and whether access to all model components is available? • How the data used for training were collected? • How the data were split into training, validation, and test sets? • What are the characteristics of the training population and its representativeness? • What model performance results were observed and what limitations were identified?
11	How will the model's performance be tested?	**Study Design**	• What performance metrics have been selected, and are they appropriate for the relevant task? Against which comparators is the AI tool evaluated in the study? • Is there a plan to distinguish between the performance of the model itself and the performance of the human–AI team?

#	Question	Category	Considerations
			• Is a staged or incremental evaluation of model performance required in order to adequately address its risk level and to support potential expansion of the study? • If algorithm updates are planned during the study, how will they affect performance measurement?
12	How will errors be identified and handled?	**Study Design**	• What are the risks associated with poor model performance? • Can errors in the model be identified and documented? • What is the user's ability to detect errors in the model output?
13	Does the tool require integration with other systems in the clinical environment?	**Data Governance**	• Where is the model hosted, is it connected to systems external to the organization, and is it integrated with internal organizational systems? • Have the risks associated with connecting to external systems or the potential impact of the model on existing organizational systems been assessed? • Is there ongoing monitoring of the organizational systems connected to the model? • Is the information required by the model expected to be available in a timely manner?
14	What does the output of the AI intervention look like? How does it contribute to the decision-making process or any other clinical activity?	**Description of the AI System**	• What is the type of output and its intended clinical purpose? • Who is the output intended for, a human user or another system? • Do users have the ability to assess the reliability of the output?

#	Question	Category	Considerations
15	Is the training process for using the AI system described in the protocol?	**Information Disclosure**	• Has it been clearly specified who is considered a user of the tool and what training requirements are required as a condition for inclusion in the study? • Does the protocol describe the need for user training for operating the system, and are the training process and covered topics clearly defined? • Has the need for usability testing or assessment of the participating teams' level of digital literacy been evaluated?
16	Is there a reference to reporting significant changes in the model to the committee? Are the criteria for early study termination well-described in the protocol?	**Study Design**	• Has the type of model been described (locked, adaptive, or continual learning)? • If the model is not locked, how are changes to the model managed during the study, and what controls are in place to govern such changes? • Has a mechanism been established for reporting material changes in the model to the ethics committee? • Have criteria for early termination of the study been defined, and under what circumstances?

Clarifications of Principles and Supplementary Guidance

1. **Does the clinical trial protocol and trial registration explicitly state that the intervention includes an AI-based component, and what is its degree of autonomy?**

   The protocol and the trial registration should explicitly state that the intervention includes an artificial intelligence–based component (e.g., AI, machine learning, or an algorithmic model). This disclosure supports transparency, ethical disclosure, and alignment with international guidance such as SPIRIT-AI, which recommends making this explicit in the title and/or abstract.

   It is also important to clarify the role of the system within the intervention and its level of autonomy or automation—whether it functions as a clinical decision support tool, a semi-autonomous system, or a system that operates independently. This distinction can help the committee assess the extent of human involvement, the potential risk level, and the type of oversight required throughout the trial.

2. **When reviewing the Informed Consent Form, does it include information on the use and degree of intervention of the AI system, and explain the purpose of use, mode of operation, risks and benefits, data handling, and reference to local regulation?**

   The informed consent form is intended to ensure that trial participants make a free and informed decision, based on an understanding of the potential risks, benefits, and available alternatives. Accordingly, when the intervention is based on artificial intelligence or machine learning, this should be clearly and explicitly stated in the consent form. The scope and manner of information disclosure should be determined in light of the participants' ability to understand the information, as well as the system's risk level, its role within the intervention, and the extent of its influence on clinical decision-making. In this context, the following elements should be considered:

   - Purpose: An explanation of how the system is used within the study (e.g., data analysis, diagnosis, treatment recommendations, or another process).

- Data handling: A description of the types of data being processed (such as imaging data, genetic information, or visit summaries), and an explanation of how data privacy and security are maintained.
- Mode of operation: A high-level description of how the system functions, including the type of model or algorithm, where relevant.
- Risks and benefits: A presentation of potential risks (e.g., errors, bias, technical failures) alongside the expected benefits (e.g., improved accuracy, increased efficiency, or enhanced quality of care).

3. What is the version of the model/tool being used?

The study protocol should explicitly specify the version of the model or tool used in the trial. Precise version identification is an important prerequisite for the reproducibility of findings, for tracking changes over time, and for ensuring appropriate regulatory and ethical oversight.

4. Has the Intended Use of the AI tool been detailed?

The study protocol should be reviewed to determine whether it provides a clear and sufficiently detailed description of the intended use of the AI tool within the clinical trial. Reference may be made to the definition of intended use as outlined in IMDRF document N81[7].

It is also important to assess whether scenarios that fall outside the intended use (out-of-scope use cases) are explicitly described, in order to reduce the risk of misuse. For example, applying the model to data types that were not included in its training may lead to unreliable or invalid outputs.

The definition of the intended use should preferably also include a description of the workflow in which the tool is expected to be integrated:

- Does the physician enter data into the system and receive a recommendation?
- In which computerized system are the results presented (e.g., an electronic medical record)?

[7] https://www.imdrf.org/sites/default/files/2024-02/IMDRFSaMD%20WGN81%20DRAFT%202024%2C%20Medical%20Device%20Software%20Considerations%20for%20Device%20and%20Risk%20Characterization%20-%20final%20draft.pdf

- Is the physician allowed to disregard or override the system's recommendation?
- Is the system used in real time during the clinical examination, or in an asynchronous manner?

In addition, any change from the standard of care should be described. It is also appropriate to clarify who the users of the outputs are (e.g., clinicians, healthcare teams, patients, or policy or payer stakeholders) and at which decision point the tool is intended to have an impact. When it is stated that output X is intended to guide decision Y, the clinical rationale and the decision pathway may be described, and, where possible, illustrated using a simple flow diagram.

When clinical decision support systems are involved, the level of human involvement in the decision should be clearly understood, and, accordingly, the system's risk level and the level of oversight required.

Finally, if the model includes an element of automatic retraining or planned updates, transparency regarding their existence and description in the protocol should be ensured. While continual training may reduce the risk of performance drift, it requires close monitoring to ensure stability and reliability. The protocol should specify how this process is monitored, who is responsible for it, and how performance assessments are adapted in such cases. Further guidance on this issue is provided in Section 15.

5. **What are the inclusion and exclusion criteria at the participant level and for the input data to the model during the study? What is the rationale for these criteria? Do these criteria match the tool's target population?**

In an interventional study protocol, the inclusion and exclusion criteria for participants should be clearly presented, along with the rationale underlying these criteria. The review committee should assess whether there is alignment between the defined criteria and the clinical indication or intended use of the tool (see Section 4), such that the selected study population appropriately represents the patient population in which the intervention is intended to

operate in practice. This alignment supports the scientific validity of the study, as discrepancies between the study population and the real-world clinical context may complicate the interpretation of results.

The relationship between the study population and the data used to train the model should also be examined (see Section 10). Substantial differences, for example, in age, ethnicity, or the devices used to collect the data, may affect model performance. Therefore, it is important to ensure that the model is applicable to the study population, or at a minimum, that there is transparency regarding differences between the training dataset and the characteristics of the inclusion population.

The following additional considerations are particularly relevant to studies based on AI tools:

- Dependence on input data: Because AI/ML tools rely on specific types of data (e.g., images, laboratory tests, clinical text), it is important to assess whether the criteria clearly define the conditions required for participant inclusion not only from a clinical perspective, but also in terms of the availability, completeness, and quality of the input data required to operate the tool.
- Integration into clinical workflows: It should be verified that the participants and data reflect the manner in which the tool is intended to be used (e.g., real-time use in an emergency department versus asynchronous use). Such alignment supports external validity.
- Fairness and representativeness considerations: The criteria should be reviewed to determine whether they allow adequate representation of relevant subpopulations (e.g., by sex, age, ethnicity, language, or socioeconomic status), and whether potential differences in model performance across populations have been considered.
- Aspects related to model updates: In cases where the model includes continual learning or periodic updates, it is important to understand whether and how this process is described, and whether it is based on data from the inclusion population defined for the study (see also Section 15).

6. **Are there prior regulatory approvals for the tool?**

   The study protocol should specify whether the product or system has previously received regulatory approval, from which regulatory authority, and whether the approval includes the artificial intelligence component. It is advisable to detail the approved indication, the version of the model that was evaluated at the time of approval (see Section 3), and whether any changes have been made to the product or the model since approval was granted. This information can assist the review committee in understanding the scope of the approved use, the existing level of regulatory oversight, and the potential need for renewed risk assessment in the event of version updates or changes in model capabilities.

7. **Is there prior information and evidence for intervention using the specific AI tool or another AI tool for the same indication/clinical purpose?**

   As part of the study review, it is important to understand what prior data or evidence exist regarding the specific AI tool for the same clinical purpose. If evidence exists relating to the same indication or to other uses that may inform the tool's safety or effectiveness, this information should be presented as part of the scientific background of the study. In addition, partial evidence, conflicting findings, or known limitations of the model (where available) should also be addressed, as these elements contribute to risk assessment and to a cautious interpretation of the study findings.

   The type of evidence presented should be clarified (e.g., retrospective studies, prospective studies, or other forms of evidence), the population in which the evidence was generated, and the performance metrics that were evaluated. In cases where comparisons were made to an existing standard (such as established clinical practice or assessments by human experts), this should be stated and the main findings briefly described.

   It is also important to understand where and under what conditions the prior evidence regarding the model was generated. The protocol should specify whether the evidence was collected within the same organization or in a clinical and operational environment similar to that in which the model will be evaluated in the current study ("internal evidence"), or in different settings ("external

evidence"). In addition, the testing conditions themselves, such as the type of data used, the model version evaluated, and the size of the dataset, should be described, in order to enable an informed assessment of the similarity between the prior evidence and the current study environment, and of the generalizability of the findings.

8. **What is the process for collecting/acquiring the data fed into the system, and how is the data processed and stored?**

   As an initial check, it should be verified that the model relies on core clinical data on which a healthcare professional would reasonably be expected to base a decision in the same context (see Section 10 on model training). Beyond this, the present section focuses on data collection for the current study, including how data are acquired, processed, and stored, as well as on potential gaps between the training environment and the real-world study environment. These aspects form the basis for examining issues related to the description of inputs, data collection configuration and methods, human involvement in the process, data quality over time, and risk management related to bias.

   The model inputs should be clearly described, including their sources (e.g., manual entry or ingestion from another system), the types of information required for operation, and the format and structure in which the data are provided. The description should specify the format in which the model expects to receive the data (e.g., standards such as FHIR or DICOM, or a proprietary structure), as well as the terminology and coding systems used (e.g., ICD, SNOMED). In the case of large language models (LLMs), the structure of the prompts, their length and content, and the specific instructions for data entry should be described. In this context, it should be taken into account that the manner in which the AI system is used—particularly the formulation and input of prompts—may have a material impact on the outputs and on the study results. Accordingly, a clear and well-documented protocol for system use should be provided, including a description of the scenarios or types of prompts to be used in the study.

   The extent of human involvement in the process should also be clarified, including whether individuals are required to enter data, validate them, or assess their

quality, and whether a certain level of expertise is required. Such involvement may affect data quality and should therefore be clearly described, including any mechanisms intended to prevent or mitigate human error.

The protocol should address how the study handles missing data, outliers, or poor-quality data, and whether there is a plan for monitoring data stability over time (data drift). Addressing these issues is part of an overall risk management approach. In addition, it is important to ensure that the data sources, collection time points, and data configurations are documented in the protocol, so that the testing conditions can be understood, oversight can be enabled if needed, and the study results can be interpreted in light of the data environment from which they were derived.

Finally, it is important to understand how issues related to data bias are examined. Bias may be introduced at different stages, including during data collection, data processing, or even during the study itself. Transparency regarding data collection methods and strategies for mitigating bias supports the validity of the study and the protection of participants.

9. **What mechanisms and tools exist for data protection, cyber security, and privacy?**

   Safeguarding privacy and ensuring data protection and cybersecurity are fundamental prerequisites for the use of artificial intelligence or machine learning tools in clinical research. The review committee should understand what privacy-preserving mechanisms exist within the organization and within the tool itself, and how these mechanisms ensure that data are not disclosed, not processed without authorization, and not used beyond their intended research purpose. Relevant information security experts within the organization should be consulted to ensure that the system complies with both technological standards and applicable regulatory and organizational requirements for protecting patient privacy, in accordance with the system's risk level.

   When third-party tools are used, such as large language models (LLMs), it is particularly important to understand how participant privacy is preserved when

data are transmitted to an external system. This includes whether the data undergo de-identification, whether it is clearly specified that the data are not used for further model training, and whether the risks of cyberattacks or data leakage have been assessed. In this context, it should be noted that the removal of direct identifiers alone does not constitute de-identification and does not provide sufficient protection against data disclosure.

Clear and explicit consideration of these aspects in the study protocol can help assess whether the protective mechanisms are appropriate given the sensitivity of the data and the risks associated with the environment in which the tool is deployed.

10. How was the model trained?

This stage of the protocol is essential for assessing the quality and validity of the model. The review should involve both AI/ML experts and clinicians, in order to ensure that the model is grounded in relevant clinical reasoning and in sound technological and scientific methods. At this stage, a comprehensive overview is required of the characteristics of the population used for training, the manner in which the data were split into training, validation, and testing sets, the data sources, and the data collection methods. It is recommended to describe how model performance was evaluated and what limitations were identified, both in terms of representativeness of subpopulations and limitations arising from data quality. These findings help clarify the degree to which the model is suitable for the target population in which it is intended to be deployed in practice.

First, it should be clarified which model (or models) are involved, who the manufacturer is and whether the manufacturer is involved in the study, whether additional models are incorporated into the process, and whether there is full access to the components of the model under evaluation.

The protocol should describe how the data used to train the model were collected, including the time point at which the data were captured (e.g., at emergency department admission, during a new examination, or in a medical summary), the devices or sources from which they were obtained, and the level of human involvement. This process can then be compared with how data are expected to

be collected in the current study. In some cases, differences between environments may be substantial, underscoring the importance of understanding potential gaps between the training data and the study data (see also Section 8).

The method used to split the data into training, validation, and testing sets, as well as the methods used to evaluate performance, should be described. It is important to understand whether the test population was held out and did not participate in the training process, whether the evaluation was temporal (based on data collected at a later time point), and whether the data underwent any preprocessing prior to use. In addition, the degree of similarity between the test data and the data used in the current study should be examined with respect to demographic, clinical, and system-level characteristics.

The characteristics of the training population should be presented, including sample size, inclusion and exclusion criteria, and key demographic and clinical features. It is important to assess whether the population is representative and whether relevant subgroups were included (e.g., by age, sex, ethnicity, or medical condition). Lack of representativeness may lead to bias and compromise the reliability and fairness of the model; therefore, known limitations and steps taken to mitigate them should be described. Consideration should also be given to less obvious but relevant subpopulations depending on the type of model, such as non-native speakers in voice-based models, rare conditions in patient summaries, or atypical presentations in image-based models.

For large language models (LLMs), a distinction should be made between the base model and any adaptations performed to enable the use of local data or knowledge sources, and the nature of these modifications and their potential impact on system performance should be described.

It is advisable to briefly present model performance as measured on dedicated test sets that were kept separate from the training process, as well as the metrics used for evaluation. The selected metrics should be appropriate to the model's intended purpose and task type (e.g., prediction, classification, text generation). In addition, it is recommended to assess whether model performance was evaluated across relevant subpopulations and whether performance differences

between groups were identified. Transparency regarding performance, metrics, and subgroup findings supports fairness assessment and helps build trust in the model.

External validation supports the assessment of the model's generalizability, particularly when the model was developed outside the organization. It should be clarified whether the model was evaluated using data collected in a clinical or organizational environment different from that of training, and how its performance compared with internal test results. If prospective performance data from real-world deployment are available, these should be briefly described (including source, population, and metrics) and their relevance to the conditions of the current study should be clarified (see also Section 7).

Finally, any limitations or issues observed during model development, such as data imbalance, variability in data quality, or risk of overfitting[8], should be described, along with any measures taken to mitigate them. It should also be considered whether the outcome prevalence in the training population reflects that of the real-world clinical population, as discrepancies may affect performance interpretation. Transparency regarding known limitations and biases strengthens the credibility of the model and the assessment of its validity in the research context.

11. How will the model's performance be tested?

At this stage, it is important to understand how the model's performance metrics have been defined and the rationale behind their selection. The review should assess whether the metrics reflect the model's intended use and the value it is designed to deliver, and whether there is a practical and reliable way to measure them.

It should be noted that the definition of performance metrics may vary depending on the nature of the task—for example, prediction, classification,

[8] In ML, overfitting occurs when a model learns the training data too thoroughly, capturing not just the fundamental patterns, but also noise or random fluctuations. Such a model might excel on the training data, but struggles to generalize to new, unseen data. https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/fda-digital-health-and-artificial-intelligence-glossary-educational-resource?utm_source=chatgpt.com#c

recommendation, segmentation, text generation, or process improvement. The investigators' consideration of what is being measured, how success will be evaluated, and what performance thresholds are expected can assist the review committee in understanding the rationale underlying the study design. Consideration may also be given to assessing output stability, including the extent to which model results are consistent in similar situations or vary in a reasonable manner in response to small changes in data or environment. It is also appropriate to examine whether performance is assessed across population subgroups, or whether there is an intention to do so at a later stage, in order to identify potential disparities or biases and to evaluate the model's fairness and consistency.

In this context, reference to performance comparisons against a benchmark (such as established clinical practice, human expert assessment, or an existing system) can help the committee understand the significance of the results: what improvements are expected (e.g., in accuracy, efficiency, speed, or novel insights), what additional risks may be introduced (such as bias or alert fatigue), how the model may affect clinical decision-making and patient outcomes, and how it integrates with existing workflows.

In addition, it is important to understand how ongoing monitoring of model performance is planned, and whether distinctions are made between the model's performance and the performance of the human–machine team[9]. The protocol should specify who is responsible for measurement, how frequently monitoring will occur, and what monitoring mechanisms will be used (e.g., dashboards, periodic sampling, or quality assessments).

Consideration should be given to whether the study design requires a staged or incremental approach to evaluating model performance, aligned with the potential risk level of the model's use. Such an approach may include, among other measures, initial deployment in a silent prospective study, limitations on the size of the population or number of users, or the definition of interim stages in an interventional (non-silent) study that include predefined stopping points and

[9] https://www.gov.il/he/pages/digital-medical-technology-gmlp-1

evaluation milestones. The existence of a clear, phased plan can assist the review committee in assessing whether sufficient risk mitigation mechanisms are in place, and how decisions regarding expansion of the study scope or increased intervention intensity are made based on observed performance.

Finally, it should be ensured that the protocol addresses potential future changes to the model, whether through planned updates or continual learning, and how such changes may affect performance during the study. In cases of substantial changes in performance or output stability, consideration should be given to the need to pause the study or to notify the IRB, in accordance with the principles outlined in Section 16.

12. How will errors be identified and handled?

The potential risks associated with suboptimal model performance (such as loss of accuracy, bias, or alert fatigue) should be assessed, and appropriate mitigation measures should be defined. Processes should be established to enable the identification and documentation of errors (for example, discrepancies between the system's output and clinical judgment). It should be examined whether mechanisms are in place to allow errors to be detected and recorded, including differences between the system's recommendations and clinicians' assessments, and the extent to which users are able to independently identify errors during system use. Such considerations can help evaluate the level of transparency, explainability, and trust in the system.

13. Does the tool require integration with other systems in the clinical environment?

In many cases, AI or machine learning tools do not operate as standalone applications but rely on integration with existing clinical systems such as electronic medical records (EMRs), databases, medical devices, imaging systems, or cloud infrastructures. The feasibility of deploying the tool may be influenced not only by its algorithmic performance, but also by its level of compatibility with the clinical environment, the availability of the data it requires, and its ability to operate reliably and safely over time. Therefore, it may be useful to assess the scope and nature of the required integration at an early stage.

It is advisable to clarify where the model is expected to operate, whether as a system installed on the healthcare organization's servers (on-premise) or as a cloud-based solution. This choice has implications for data protection, cybersecurity, and privacy mechanisms, including access control, encryption, access monitoring, and compliance with relevant regulatory standards.

It is important to understand how frequently and in what manner data are fed into the model, and how the availability of the required data at the appropriate time is ensured. Consideration may be given to whether the model operates in batch mode, is triggered by clinical events (event-based), or relies on continuous real-time data streams, and whether the information available in organizational systems is accessible at the temporal resolution required for this purpose.

Beyond establishing the integration, consideration should be given to how it will be maintained over time. Changes in source systems, software updates, or interface modifications may affect the tool's functionality. Addressing issues such as fault monitoring, support mechanisms, and response procedures can help ensure operational continuity and system stability.

When assessing integration, it should be taken into account that connecting a new tool to the clinical environment may expose the organization to information security risks, data leakage, or disruption to the continuity of existing systems. Two primary risk directions should be considered: on the one hand, connecting the model to external systems may expose the organization to intrusions, data leaks, or dependencies on external entities (such as cloud services or technology vendors); on the other hand, the model itself may impact organizational systems, for example by modifying or writing to existing data, consuming significant resources, or storing sensitive information in an uncontrolled manner. See also Section 9.

14. What does the output of the AI intervention look like? How does it contribute to the decision-making process or any other clinical activity?

It is important to understand the intended purpose of the output and the type of information produced by the AI system. The output may be intended to provide clinical interpretation or intervention (such as a diagnosis, risk stratification, or

treatment recommendation), to support a workflow process (such as referral for additional testing or selection of technical parameters), or to generate data for clinical processing (such as measurements, segmentation, or processed images). The type of output variable should also be clarified (e.g., binary, numerical, textual, visual, or other). In cases of textual output, particularly with large language models (LLMs), it may be useful to understand the structure of the output, whether it relies on existing documents (as in retrieval-augmented generation models), and the characteristics of the sources on which it is based.

It should be clarified for whom the output is intended—whether it is presented to a human user (such as a physician, nurse, technician, or patient) or transmitted to another medical device—and the level of autonomy of the system should be examined. This includes whether the system operates under human oversight or functions as a clinical decision support tool in which the final decision remains with the human user.

It should be clarified for whom the output is intended—whether it is presented to a human user (such as a physician, nurse, technician, or patient) or transmitted to another medical device—and the level of autonomy of the system should be examined. This includes whether the system operates under human oversight or functions as a clinical decision support tool in which the final decision remains with the human user.

Attention should also be paid to the user's ability to detect errors or deviations in the output. In cases where the output is visual—for example, highlighting a suspected region—the potential for human oversight may be relatively high; by contrast, in long or complex textual outputs, it may be more difficult to identify subtle errors that could affect clinical meaning. In such situations, it may be appropriate to consider assessing the level of difficulty in error detection and its potential impact on study outcomes. See also Section 11.

15. Is the training process for using the AI system described in the protocol?

Because the user is an integral part of the intervention (e.g., physician, nurse, or patient), it is important to understand who is authorized to use the system within

the study, how the system is intended to be used by them, and what training and oversight measures have been established. The protocol should be reviewed to determine how the user training process for the AI system is described, and whether the involved teams receive dedicated training prior to or during the study. Addressing the components of training can help clarify how appropriate and cautious use of the system is ensured, including how users are informed about the study objectives, the proper use of the system (including, where relevant, expanded guidance on prompts—see Section 8), and the system's potential limitations.

It is also appropriate to consider whether a certain level of digital literacy or technological knowledge is required for participation in the study, and how such knowledge is assessed or acquired in practice. In addition, consideration may be given to whether the training process or the conduct of the study includes an assessment of usability aspects, such as interface clarity, comprehension of instructions, and users' ability to interact with the system in a consistent and safe manner.

In cases involving clinical decision support systems, the training process is of particular importance. Study outcomes may be significantly influenced by the quality of collaboration between humans and the AI system (the human–AI team). Training should ideally support users in understanding when and how it is appropriate to rely on system outputs, how human oversight of system actions is maintained, and how algorithmic recommendations are integrated into the clinical workflow in a thoughtful and cautious manner.

16. **Is there a reference to reporting significant changes in the model to the committee? Are the criteria for early study termination well-described in the protocol?**

The model undergoing changes during the study may be a locked model, a continuously learning model, or a model that is updated in a planned manner. It is advisable to clarify to the committee which type of model is involved and what controls the investigators have over changes to its components. The committee may request details regarding the frequency of updates, monitoring mechanisms,

and control measures intended to ensure that updates do not compromise participant safety or the reliability of study results. In addition, it should be clarified whether the model relies on external models or third-party components, and whether the research team has transparency and control with respect to changes made to these components.

Because changes to the model may affect its performance, it is important to assess how such changes may influence the conduct of the study, its outcomes, and the associated risk management. In cases where changes may be substantial, consideration may be given to establishing a mechanism for reporting or renewed review by the committee. Changes in model performance resulting from updates or modifications during the study may justify reassessment of study continuation; therefore, it is recommended to define critical thresholds that will be evaluated in a planned manner throughout the study.

In addition, the protocol may specify additional conditions under which early termination of the study would be considered, whether due to safety concerns, lack of effectiveness, or concerns regarding unexpected effects on participants or on the reliability of the results. Situations that may justify consideration of early termination include, but are not limited to:

- Serious or unusual safety events indicating risk to patients.
- Repeated failures in integration or operation of the AI system.
- Ongoing difficulty among clinical staff in interpreting or using system outputs (a structured reporting mechanism may be considered, through which clinicians document problematic outputs for real-time monitoring).
- Input data that are insufficient to enable consistent operation of the system.
- Abnormal or unexpected outputs identified by the system during use.
- Significant changes in the clinical environment (such as changes in equipment, data structures, treatment guidelines, or drug formularies) that may materially affect model performance.

The committee may also consider early termination in the opposite scenario, where data indicate a clear and substantial success of the intervention. In such

cases, continuation of the study may be unnecessary and may raise ethical concerns if it delays patient or system access to an effective and safe intervention.

It is advisable for the protocol to address factors that may lead to substantial changes in the model, how such changes will be evaluated, the frequency of review, and the circumstances under which an update or report to the committee will be considered.

Types of studies across the AI lifecycle

There is sometimes a lack of clarity regarding the distinction between different types of studies based on artificial intelligence (AI), including when a study is considered interventional, when it is observational or silent, and the role of each type across the lifecycle of development and implementation of the tool. For this purpose, the following section presents a classification into four main types of studies, integrated across the different stages of development, training, evaluation, and deployment of AI tools in the healthcare system.

This clarification is also intended to emphasize that the present document focuses on the review of prospective interventional ("active") studies, in which the use of the tool may have an actual impact on the clinical decision-making process.

- **Studies related to model development and training, followed by retrospective evaluation of the performance of models that have not yet been implemented:** This stage includes non-interventional studies whose objective is to train and/or evaluate different models and to identify the model with the best performance using retrospective data. In such cases, it is important to take into account aspects of data protection, cybersecurity, and patient privacy, in accordance with applicable local regulations and ethical requirements.

  Example: A study of a system for early detection of signs of myocardial infarction based on ECG data. At this stage, several different models are developed or evaluated using existing ECG data in order to identify the model with the best performance in predicting myocardial infarction.

- **Prospective silent study:** This stage includes non-interventional studies whose objective is to evaluate the model in real-world settings without affecting patients. If the study requires integration of the model into the operational systems of the healthcare organization, additional considerations arise regarding potential impacts on the functioning of organizational systems. Such integration introduces further challenges related to information security and ongoing de-identification of data, which require attention in accordance with applicable local regulatory and ethical requirements.

***Example**: A study of a system for early detection of signs of myocardial infarction that is operated in silent mode in hospitals, where it analyzes ECG data in real time but does not influence treatment decisions. At this stage, a comparison is made between the relevant department's records of myocardial infarction events and the system's predictions, in order to evaluate its performance.*

- **Prospective interventional ("active") study:** This stage includes studies whose objective is to assess the impact of models within an interventional trial in real-world settings, in a manner that may affect the course of treatment. The present document addresses situations of this type, and ethical and safety issues related to the unique characteristics of AI/ML technologies, as presented in the table in this document, should be examined.

  ***Example**: A study of a system for early detection of signs of myocardial infarction that is operated in active mode and provides real-time alerts to the medical staff regarding patients at high risk of myocardial infarction. At this stage, the effects of the system on treatment decisions, the system's accuracy in preventing myocardial infarctions, and additional ethical and safety challenges are evaluated.*

- **Post-implementation evaluation study:** This stage includes studies whose objective is to assess the performance of models that are already in routine clinical use. In these cases as well, it is important to consider aspects of data protection, cybersecurity, and patient privacy, in a manner similar to the principles and considerations applied to retrospective studies and prospective silent studies.

  ***Example**: A study evaluating the effectiveness of the implementation of a system for early detection of fractures in medical imaging, which is operated in the hospital as part of routine clinical practice and alerts medical staff to possible findings of current or future disease.*

Terminology

This section brings together basic concepts required for understanding the document. Its purpose is to make the key terms accessible to readers and to ensure a shared language among the various stakeholders. The explanations below are presented concisely and are intended to facilitate reading and practical application of the document. For readers interested in further exploration, references are provided to additional sources and expanded materials that include further terminology from the field.

Artificial Intelligence (AI) is a branch of computer science, statistics, and engineering that uses algorithms or models to perform tasks and exhibit behaviors such as learning, making decisions, and making predictions[10].

Machine Learning (ML) - Machine learning models are a subset of artificial intelligence (AI) models in which the computer learns from data rather than being explicitly programmed. Learning may be supervised (where the system is provided with labeled correct answers), unsupervised (where the system independently identifies patterns or groupings in the data), or semi-supervised (a combination of the two). Additional approaches also exist, such as reinforcement learning, in which the system learns through trial and error based on feedback or rewards.

Generative Artificial Intelligence (Generative AI) - The class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content.

This is usually done by approximating the statistical distribution of the input data. For example, in healthcare, generative AI can be used to generate annotations on synthetic medical data (e.g., image features, text labels) to help expand datasets for training algorithms[11].

Large Language Model (LLM) - A type of AI model trained on large-scale text corpora with the aim of learning relationships between words in natural language. These

[10] https://www.imdrf.org/sites/default/files/2022-05/IMDRF%20AIMD%20WG%20Final%20Document%20N67.pdf
[11] https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/fda-digital-health-and-artificial-intelligence-glossary-educational-resource?utm_source=chatgpt.com#c

models apply the patterns they have learned to predict and generate natural language responses to a wide range of inputs or prompts, and to perform tasks such as translation, summarization, and question answering. They are characterized by a very large number of parameters (internal variables learned during the training process).

Locked Model - a model that provides the same output each time the same input is applied to it and does not change with use, as its parameters or configuration cannot be updated. In case of AI-enabled medical products, locked models can help ensure consistent performance[12].

Continual Machine Learning - The capability of a model (whether locked or adaptive) to improve its performance over time by incorporating new data or experience, while retaining previously acquired knowledge and information. Model updates are implemented in such a way that, for a given set of inputs, the output may differ before and after the change. Such changes are typically carried out and validated through a well-defined and structured process aimed at improving model performance based on analysis of newly available data.

Retrieval-Augmented Generation (RAG) is a technique developed to improve how LLMs, such as those behind advanced chatbots and virtual assistants, handle information. For different reasons, including reliance on old data, LLMs can provide incorrect answers, and it can be difficult to understand how they arrived at a particular response. RAG allows LLMs to access additional data sources that can keep information current, which is particularly useful when applied to specialised domains or knowledge areas[13].

The degree of autonomy is a spectrum of capacities or liberties to operate independently of a user's direction, intervention and oversight. An autonomous device or medical device software provides outputs that impact the subsequent clinical action or decision without any user intervention. Conditionally autonomous outputs will meet this condition selectively (for example, for certain results, input

[12] https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/fda-digital-health-and-artificial-intelligence-glossary-educational-resource#o
[13] https://www.oecd.org/en/publications/governing-with-artificial-intelligence_795de142-en/full-report.html

characteristics, or circumstances). Supervised outputs can impact subsequent clinical actions or decisions without a user having to approve the output but operate under the supervision of users. Non-autonomous outputs are typically intended to help a user during their determination of a decision or action. The degree of clinical autonomy can be made clear by describing this aspect for clinical users specifically [14]. **It should be emphasized that this document does not address autonomous systems, for which the assessment of additional aspects is required and is beyond the scope of this document.**

Clinical decision support (CDS) is a software function that provides health care professionals and patients with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care[15].

[14] https://www.imdrf.org/sites/default/files/2025-01/IMDRF_SaMD%20WG_Software-Specific%20Risk_N81%20Final_0.pdf?utm_source=chatgpt.com
[15] https://www.fda.gov/medical-devices/digital-health-center-excellence/step-6-software-function-intended-provide-clinical-decision-support?utm_source=chatgpt.com