IMDRF International Medical Device Regulators Forum | EU2023 EUROPEAN UNION Chair

European Commission | European Union

AI specific post-market clinical follow-up endpoints

- **Leo Hovestadt (DITTA CIE WG Chair)**

- **13 March 2023**

Overview – AI endpoint categories

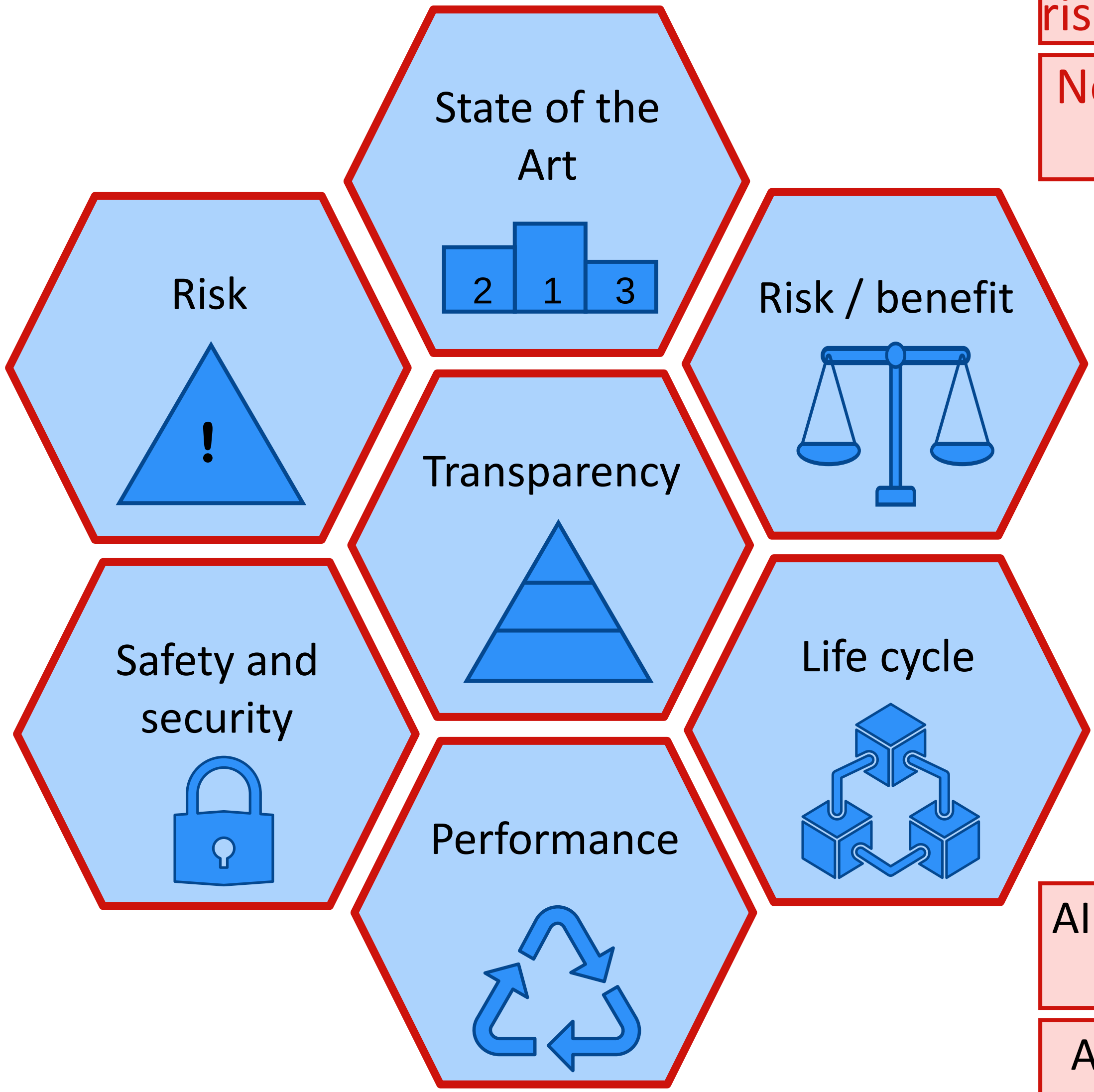
Agenda:

AI endpoints categories

Transparency matrix and endpoints

Insider transparency: PMCF endpoints

Internal transparency: IDEAL endpoints



Not one definition of risk (SaMD, AI, Ethics, …)

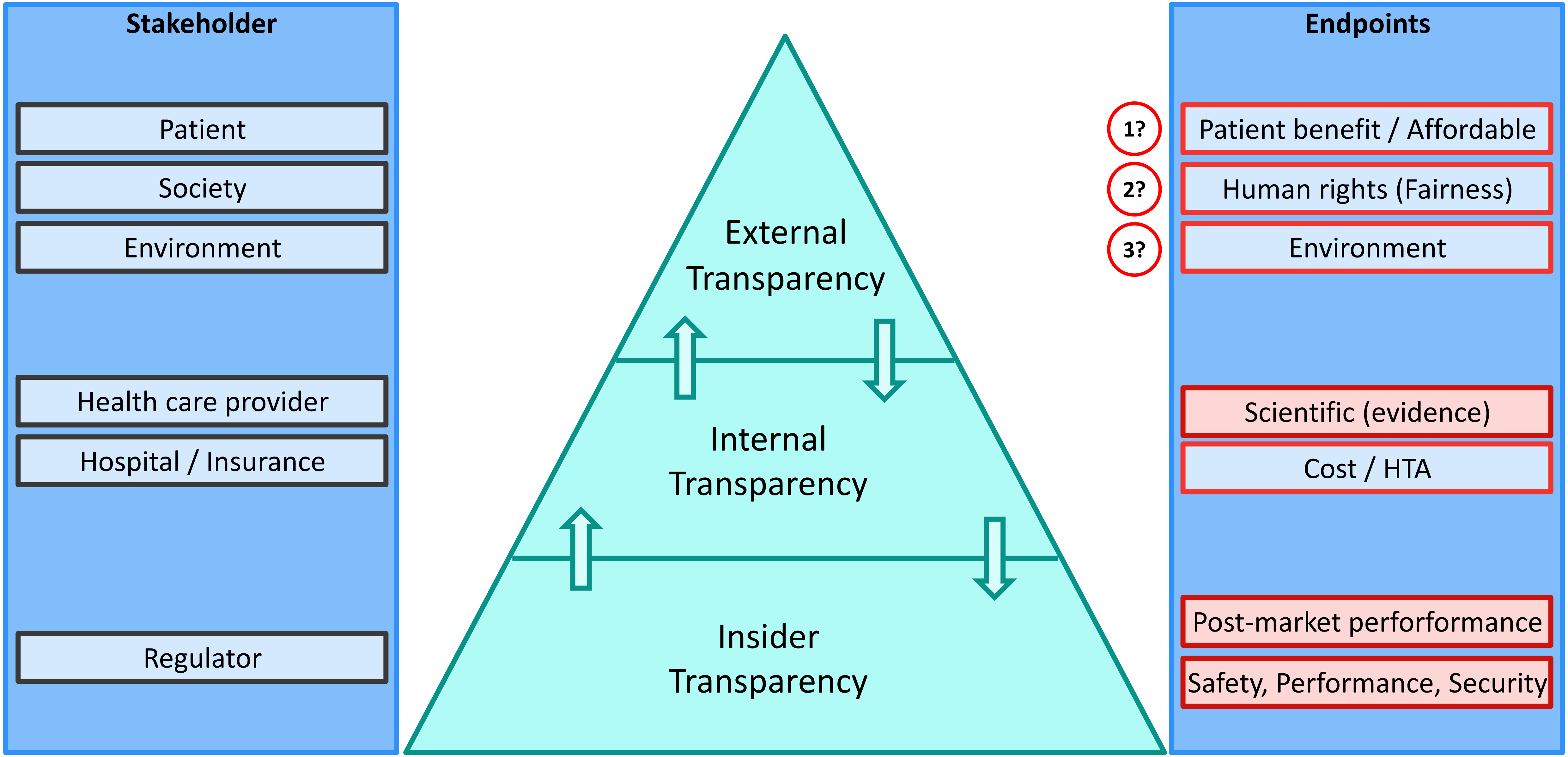
Not all AI risks weighed against benefits

Regulator help is appreciated to keep regulations and standards medical device specific

AI state of the art changes quickly

AI performance changes over lifetime

Transparency matrix and AI endpoint stakeholders



The transparency matrix explains easily this AI concept

Insider transparency: AI PMCF endpoints and SamD life cycle and GMLP

IEC 62304 Software development life-cycle processes & GMLP

Global System Requirements

SW Requirements

SW Design Process

SW Code Phase

Data Management GMLP 4, 5	**Post market follow up** GMLP 7, 10
Configuration Management	Change Management
Quality Assurance Process GMLP 1, 2	Planning Process: GMLP 6, 9
Risk Management GMLP 1, 2, 6, 10	Test Management GMLP 3, 8

Design Validation SW / HW

SW Verification

Unit Test Phase

Code Review Phase

GMLP PMCF endpoint examples

GMLP 6, 10 Commissioning: Endpoint for overfitting of deployed model

GMLP 6 Life cycle: Endpoint for performance degradation

GMLP 7 Life cycle: Endpoint for growing usability issues

PMCF example related issues

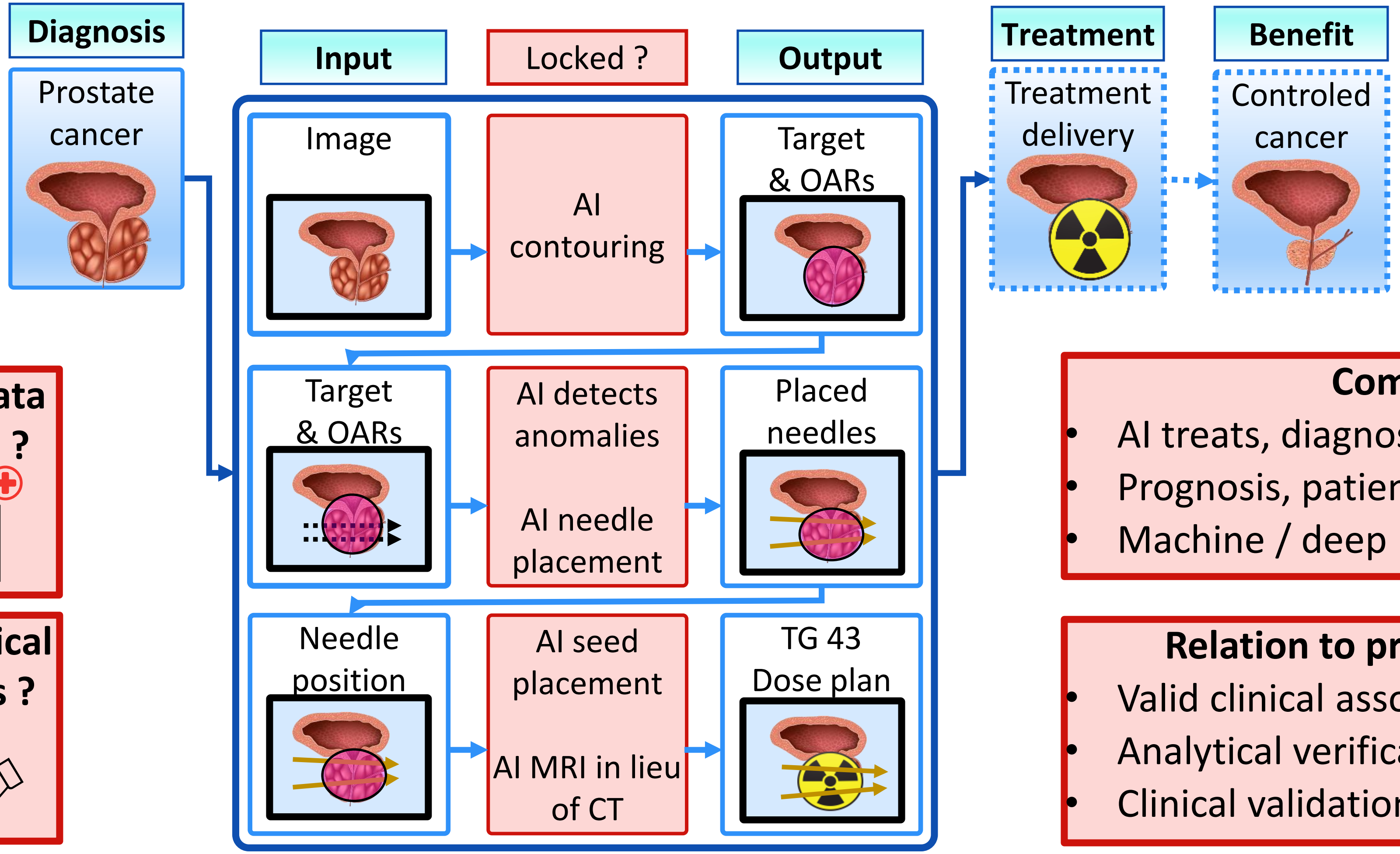
GMLP 10 Life cycle: Plan algorithm retraining and change management

GMLP 10 Life cycle: Plan usage of sandboxes

GMLP 10 End of life: Manage disposal of the training data

GMLP integrates well in IEC 62304

Insider transparency AI PMCF endpoint reflections



“Clinical” data definitions ?

Limited clinical data access ?

Complexities ?

- AI treats, diagnosis, drives, informs, …
- Prognosis, patient management…
- Machine / deep learning, unsupervised, …

Relation to premarket endpoints?

- Valid clinical association
- Analytical verification / validation
- Clinical validation

PMCF vs ISO 14155 ?

GMLP integrates well in IEC 62304

Internal transparency: R-IDEAL endpoints for radiotherapy: reflections

Milestones	Purpose	Endpoints (Outcomes)	Study design
Stage 0 Predicate studies	• How to use the innovation (software, coils needed)? • Why and in whom to use the innovation?	• MR sequences, dedicated coils, etc. • Inter-rater reproducibility • Treatment strategies, patient selection	Phantom studies, delineation studies, planning studies, model-based studies
Stage 1 Idea	First time use of the innovation for treatment delivery in men	Proof of concept	Structured case report
Stage 2a Development	Technical optimization of the innovation for treatment delivery	Technical improvements, feasibility, and safety	Prospective small uninterrupted case series
Stage 2b Exploration	Provide proof of early clinical effectiveness and safety of the innovation	Early effectiveness: • toxicity • tumor response • local recurrence (with spacious information)	Prospective study with preferably randomized component: RCT; cmRCT; random allocation of limited available treatment slots to eligible patients; Comparison with matched (historical) controls
Stage 3 Assessment	Formal comparison of innovation against standard treatment Development of clinical guidelines?	Effectiveness compared to standard treatment: • (disease-free) survival /recurrence / toxicity • PROMs, CTC-PRO, Cost effectiveness	RCT, cmRCT, registry-based trial Scientific (evidence)? Safety, Performance
Stage 4 Long-term evaluation	Long-term outcomes of the innovation, post-marketing, and surveillance Clinical guidelines by clinicians or manufacturers ?	Long-term toxicity, long-term (disease-free) survival, rare side effects, Patient-Reported Outcomes	Prospective registries, including all patients treated with the innovation Post-market monitoring

THANK YOU

Contact information:

- Leo.Hovestadt@elekta.com

Back-up slide - Good Machine Learning Practices (GMLP)

1. Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle: In-depth understanding of a model's intended integration into clinical workflow, and the desired benefits and associated patient risks, can help ensure that ML-enabled medical devices are safe and effective and address clinically meaningful needs over the lifecycle of the device.

2. Good Software Engineering and Security Practices Are Implemented: Model design is implemented with attention to the "fundamentals": good software engineering practices, data quality assurance, data management, and robust cybersecurity practices. These practices include methodical risk management and design process that can appropriately capture and communicate design, implementation, and risk management decisions and rationale, as well as ensure data authenticity and integrity.

3. Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population: Data collection protocols should ensure that the relevant characteristics of the intended patient population (for example, in terms of age, gender, sex, race, and ethnicity), use, and measurement inputs are sufficiently represented in a sample of adequate size in the clinical study and training and test datasets, so that results can be reasonably generalized to the population of interest. This is important to manage any bias, promote appropriate and generalizable performance across the intended patient population, assess usability, and identify circumstances where the model may underperform.

4. Training Data Sets Are Independent of Test Sets: Training and test datasets are selected and maintained to be appropriately independent of one another. All potential sources of dependence, including patient, data acquisition, and site factors, are considered and addressed to assure independence.

5. Selected Reference Datasets Are Based Upon Best Available Methods: Accepted, best available methods for developing a reference dataset (that is, a reference standard) ensure that clinically relevant and well characterized data are collected and the limitations of the reference are understood. If available, accepted reference datasets in model development and testing that promote and demonstrate model robustness and generalizability across the intended patient population are used.

6. Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device: Model design is suited to the available data and supports the active mitigation of known risks, like overfitting, performance degradation, and security risks. The clinical benefits and risks related to the product are well understood, used to derive clinically meaningful performance goals for testing, and support that the product can safely and effectively achieve its intended use. Considerations include the impact of both global and local performance and uncertainty/variability in the device inputs, outputs, intended patient populations, and clinical use conditions.

7. Focus Is Placed on the Performance of the Human-AI Team: Where the model has a "human in the loop," human factors considerations and the human interpretability of the model outputs are addressed with emphasis on the performance of the Human-AI team, rather than just the performance of the model in isolation.

8. Testing Demonstrates Device Performance during Clinically Relevant Conditions: Statistically sound test plans are developed and executed to generate clinically relevant device performance information independently of the training data set. Considerations include the intended patient population, important subgroups, clinical environment and use by the Human-AI team, measurement inputs, and potential confounding factors.

9. Users Are Provided Clear, Essential Information: Users are provided ready access to clear, contextually relevant information that is appropriate for the intended audience (such as health care providers or patients) including: the product's intended use and indications for use, performance of the model for appropriate subgroups, characteristics of the data used to train and test the model, acceptable inputs, known limitations, user interface interpretation, and clinical workflow integration of the model. Users are also made aware of device modifications and updates from real-world performance monitoring, the basis for decision-making when available, and a means to communicate product concerns to the developer.

10. Deployed Models Are Monitored for Performance and Re-training Risks are Managed: Deployed models have the capability to be monitored in "real world" use with a focus on maintained or improved safety and performance. Additionally, when models are periodically or continually trained after deployment, there are appropriate controls in place to manage risks of overfitting, unintended bias, or degradation of the model (for example, dataset drift) that may impact the safety and performance of the model as it is used by the Human-AI team.