

Regulatory Science Challenges for Generative AI Applications in Medical Devices

Digital Health Advisory Committee

November 20, 2024

Victor Garcia and Aldo Badano

Office of Science and Engineering Laboratories

Center for Devices and Radiological Health

U.S. Food and Drug Administration

Regulatory science for accelerating patient access to innovative, safe and effective medical devices



Office of Science and Engineering Labs (OSEL/CDRH/FDA)

Dedicated to promoting innovation for the development of new lifesaving medical devices

OSEL is organized into 20 program areas

AI/ML program is one of the largest

OSEL outputs are regulatory science tools (RSTs)

Innovative tools for assessing safety or effectiveness of emerging technology that innovators can readily (and voluntarily) incorporate into all stages of device development

An official website of the United States government [Here's how you know](#) ▼

FDA U.S. FOOD & DRUG ADMINISTRATION

Search [] Menu []

Regulatory Science Tools Catalog

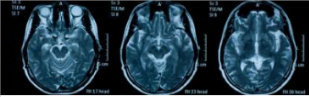

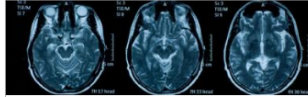

Search Tool Catalog [] Search []

Tools Categories

- Lab Method (30)
- Computer Model (21)
- Dataset (6)
- Phantom (2)
- Physical (1)
- Clinical Outcome Assessment (1)

Program Areas

- Cardiovascular (18)
- Medical Imaging and Diagnostics (13)
- Orthopedic Devices (8)
- Biocompatibility and Toxicology (6)
- Credibility of Computational Models (6)

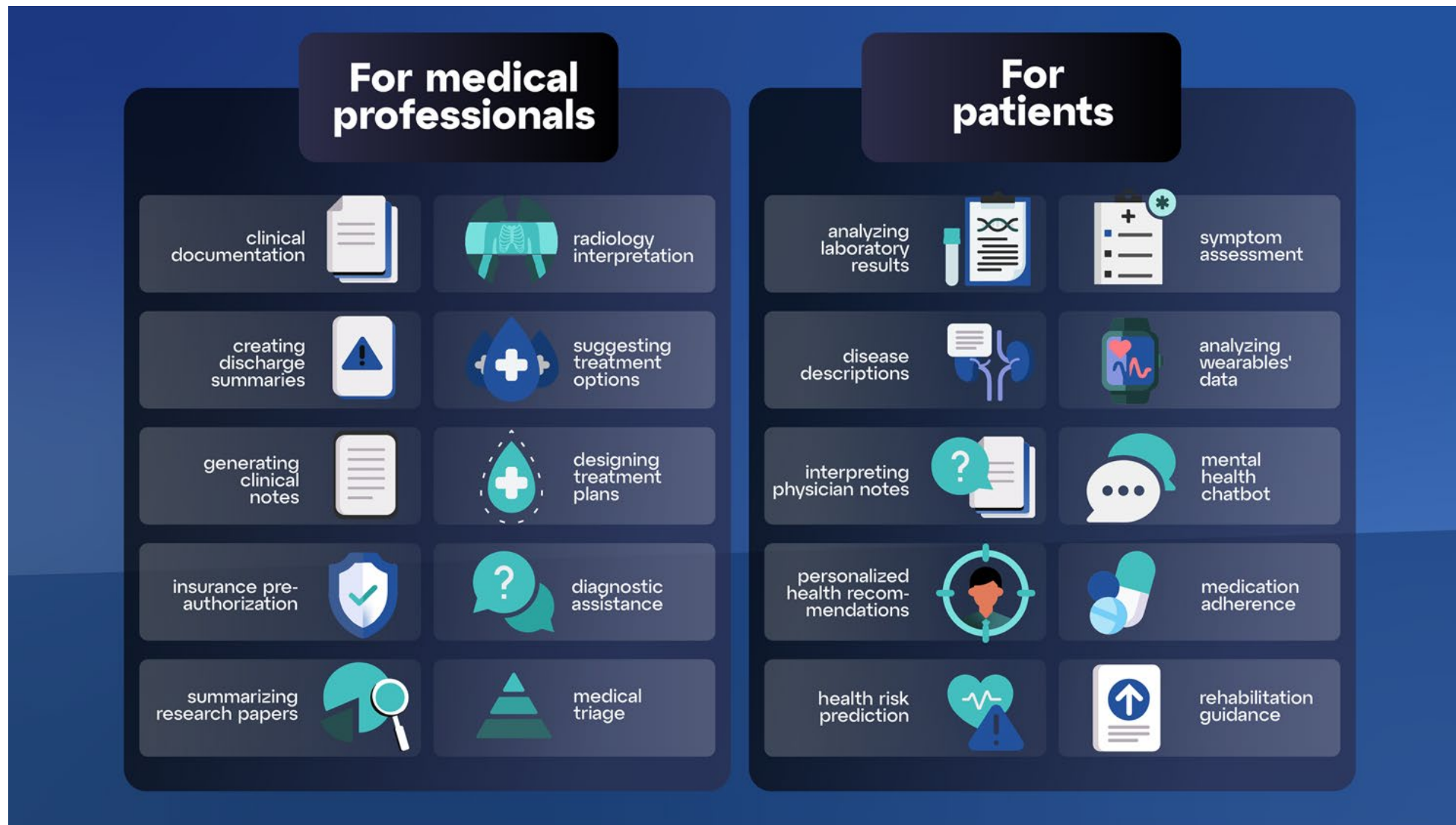
 <p>VICTRE: In Silico Breast Imaging Pipeline</p> <p>Computer Model</p> <p>The Virtual Imaging Clinical Trials for Regulatory Evaluation (VICTRE) computer modeling pipeline is a set of tools that allow for the replication of...</p> <p>Medical Imaging and Diagnostics</p>	 <p>Targeted Box and Blocks Test (tBBT)</p> <p>Clinical Outcome Assessment</p> <p>A performance-based method requiring controlled grasping, transport, and release of objects that can be used to evaluate upper limb functional ability.</p> <p>Human Device Interaction Orthopedic Devices Neurology</p>	 <p>The Virtual Family: A set of anatomically correct whole-body computational models</p> <p>Computer Model</p> <p>The Virtual Family provides detailed three-dimensional computational models of the human anatomy including an adult male, an adult female, and tw...</p> <p>Orthopedic Devices Ophthalmology Neurology Medical Imaging and Diagnostics ...</p>	 <p>Toolkit for Evaluation of Head Mounted Display Image Quality</p> <p>Lab Method</p> <p>This tool allows for the creation of immersive 3D scenes using a web browser. WebXR allows for an instant deployment of any 3D scene and script...</p> <p>Medical Extended Reality</p>
--	--	--	--

www.fda.gov/about-fda/cdrh-offices/office-science-and-engineering-laboratories

Contact at OSEL_CDRH@fda.hhs.gov
RST Catalog: <https://cdrh-rst.fda.gov/>



Generative AI in Healthcare



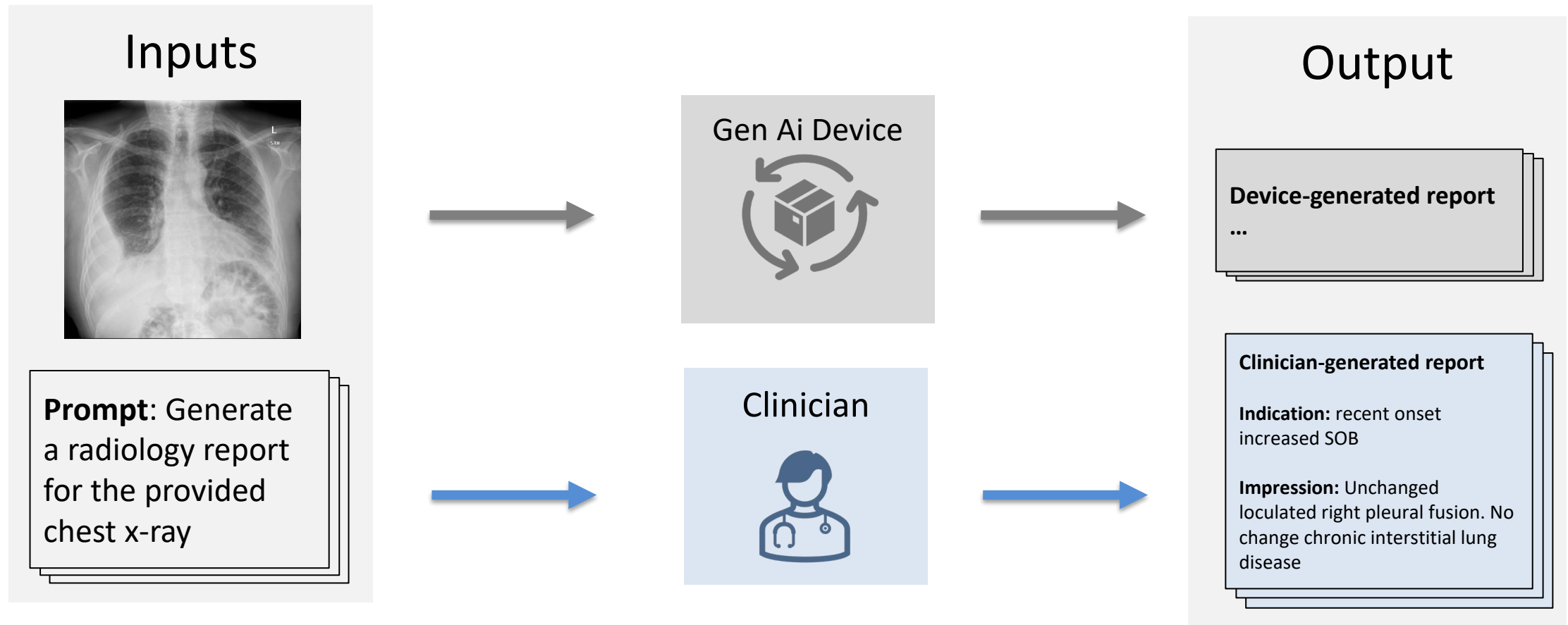
Meskó, B., Topol, E.J. The imperative for regulatory oversight of LLMs (or generative AI) language models in healthcare. *npj Digit. Med.* 6, 120 (2023)

Regulatory Science Challenges for GenAI-Enabled Medical Devices



- Difficulty in defining scope of product's intended use, e.g., due to open-ended inputs and outputs
- Foundation models not under the provenance of medical device manufacturer
- Providing oversight for an adaptive system
- Identification of hallucinations
- Adequacy of data sets for testing, including diversity
- Evaluation of and monitoring for performance in the real world, including bias
- Providing transparency to users

Example Use Case of Generative AI in Radiology



Source: [Indiana University Chest X-ray Collection](#) | [Open-i](#)
©Copyright Policy- open-access [License](#)
No changes were made.

Performance Assessment Strategies for GenAI in Healthcare



Benchmarking

with standardized
reference datasets

Expert evaluation

including holistic
evaluation of AI-
human interactions

Model-based evaluation

including automated
testing

Performance Assessment Strategies: **BENCHMARKING**



What is it?

Evaluating models on specific tasks using external test datasets and predetermined metrics.

Advantages

- Practical and available
- Allows head-to-head comparisons
- Large scale

Disadvantages

- Limited in tasks and datasets
- Train-to-the-test overfitting

The screenshot shows the LLM Benchmark interface with the following sections:

- LLM Benchmark** (header)
- Submit** (button)
- Model Vote** (button)
- Search**: Input field with placeholder "Separate multiple queries with ';"
- Select Columns to Display**: Grid of checkboxes for metrics and attributes such as Average, IFEval, BBH, MATH Lvl 5, GPQA, MMLU-PRO, CO₂ cost (kg), etc.
- Model types**: Filtered categories like chat models, fine-tuned on domain-specific datasets, base merges and moerges, pretrained, multimodal, and continuously pretrained.
- Precision**: Options for bfloat16, float16, and 4bit.
- Select the number of parameters (B)**: Slider and input fields for 7 and 10.
- Hide models**: Filtered options like Deleted/incomplete, Merge/MoErge, MoE, and Flagged.



Performance Assessment Strategies: **EXPERT EVALUATION**

What is it?

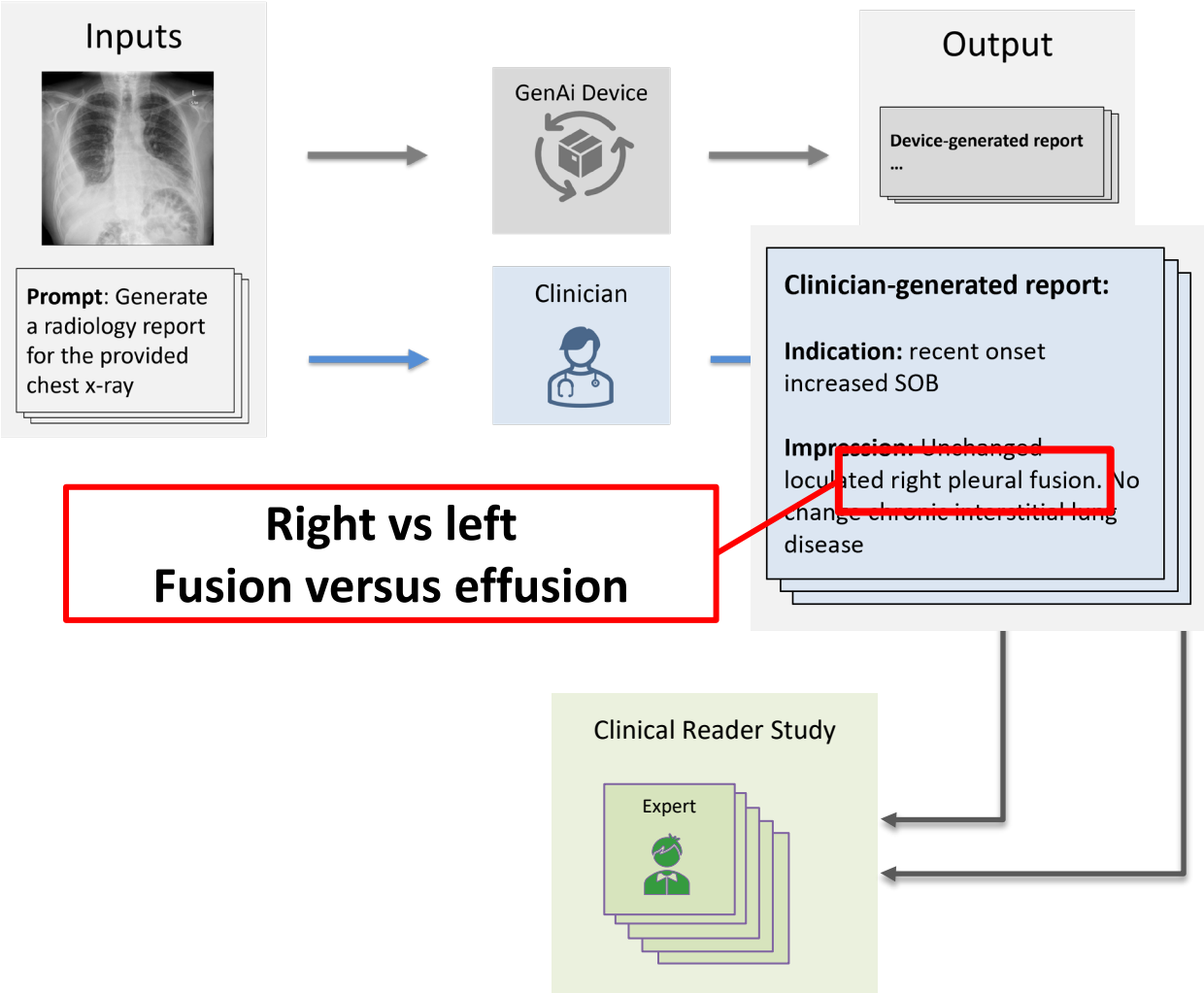
Evaluating models using expert annotations as the reference standard.

Advantages

- Adaptable to new medical tasks
- Direct clinical relevancy

Disadvantages

- Resource intensive
- Subjective and highly variable



Performance Assessment Strategies: MODEL-BASED EVALUATION



What is it?

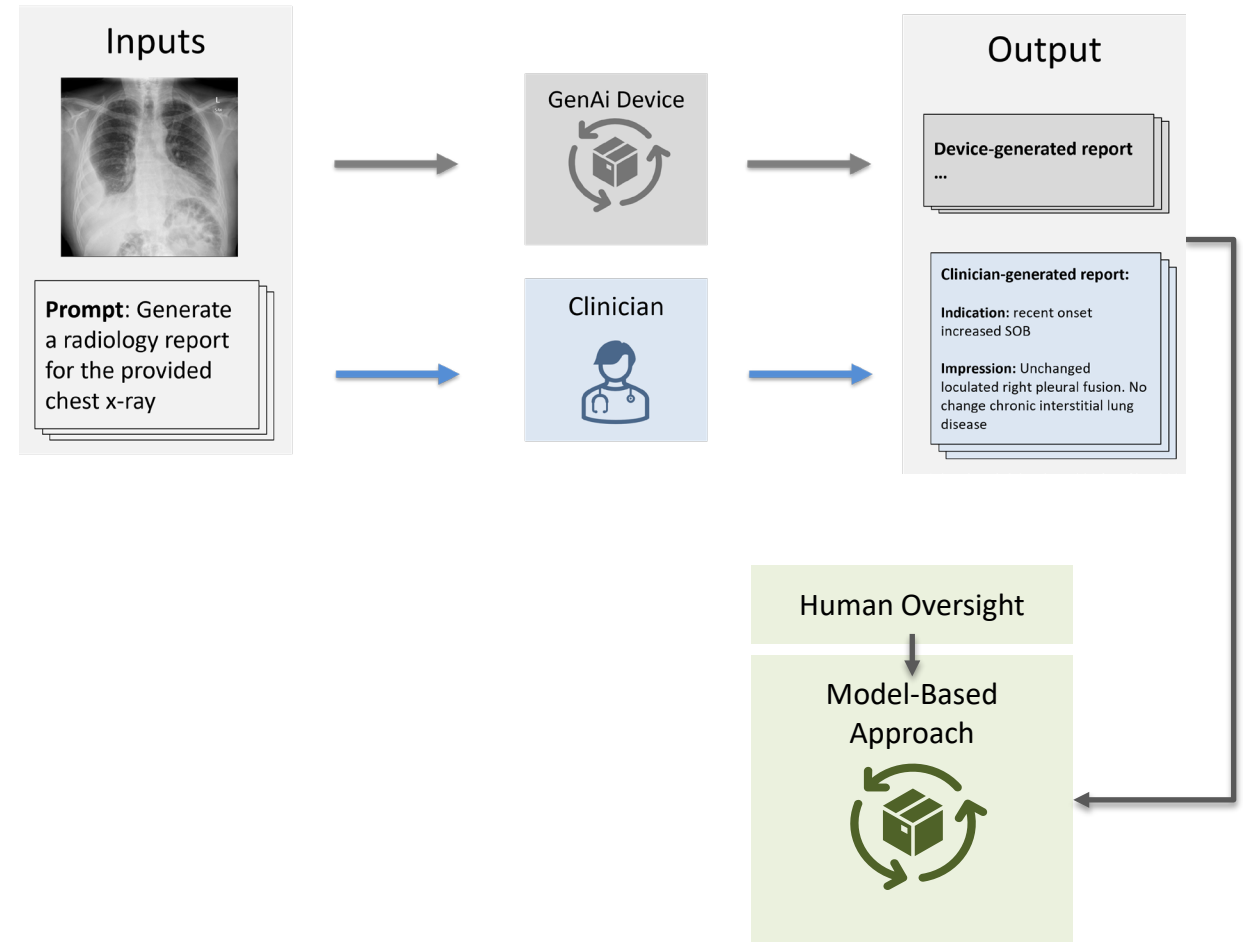
Evaluating models using a *model-based approach* (may be based on genAI) with human oversight

Advantages

- Augments human evaluation
- Scalable

Disadvantages

- Burdensome validation
- Inter-model leakage



Performance Assessment Strategies: MODEL-BASED EVALUATION



What is it?

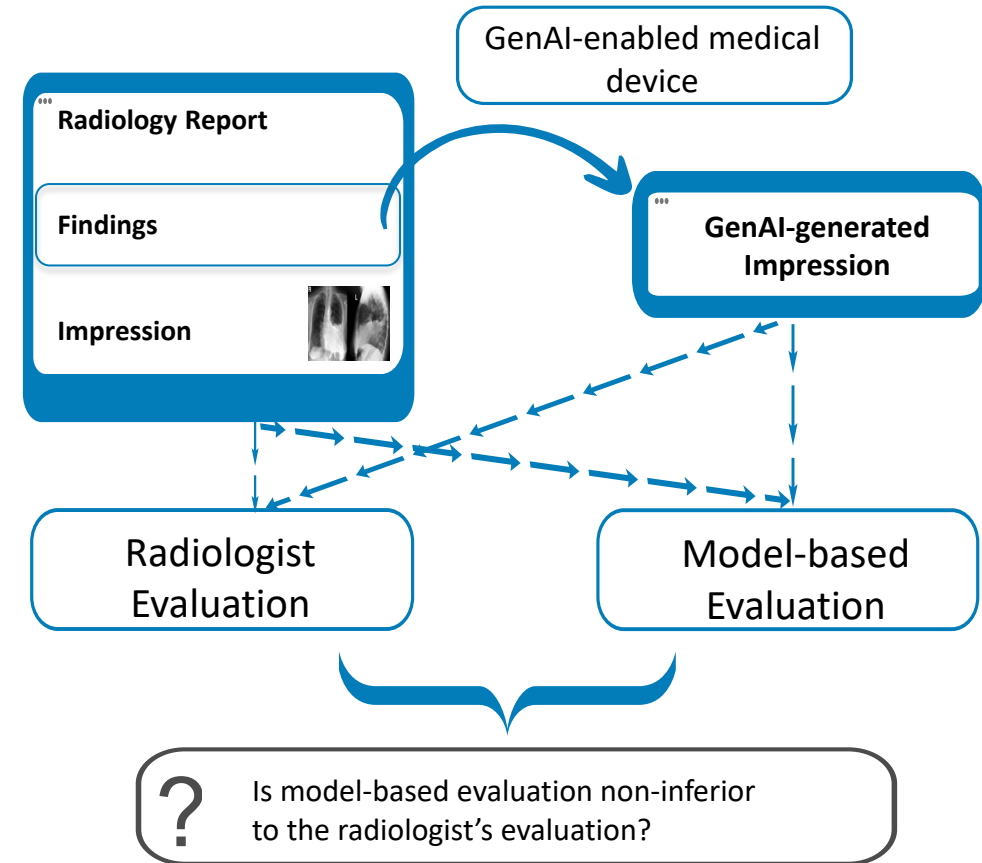
Evaluating models using a *model-based approach* (may be based on genAI) with human oversight

Advantages

- Augments human evaluation
- Scalable

Disadvantages

- Burdensome validation
- Inter-model leakage



Current research in OSEL aims at developing a case-agnostic approach to characterizing factual accuracy: e.g., are the findings in the genAI-generated report found in the reference report?

Summary



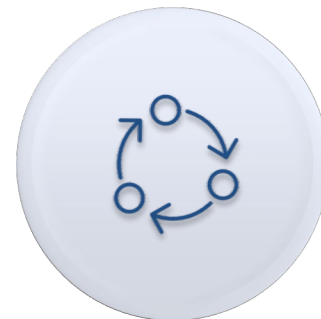
For some GenAI, known evaluation strategies may apply



For some GenAI, new evaluation methodologies and new performance metrics may need to be developed



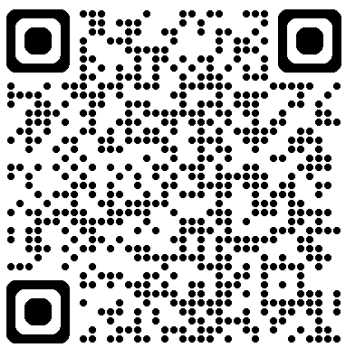
Performance evaluation requirements are governed by the intended use and associated risk



Totality of evidence may include pre- and post-market elements

Thank you for your attention

Victor Garcia victor.garcia@fda.hhs.gov
Aldo Badano aldo.badano@fda.hhs.gov
Office of Science and Engineering Laboratories
Center for Devices and Radiological Health
U.S. Food and Drug Administration



▶ www.fda.gov/about-fda/cdrh-offices/office-science-and-engineering-laboratories