

Healthcare AI: Challenges in Regulatory Science

National Artificial Intelligence Research Resource (NAIRR) Software
Workshop: A comprehensive and accessible AI software stack

December 3, 2024

Aldo Badano

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  [aldobadano](#)  [aldobadano](#)

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)

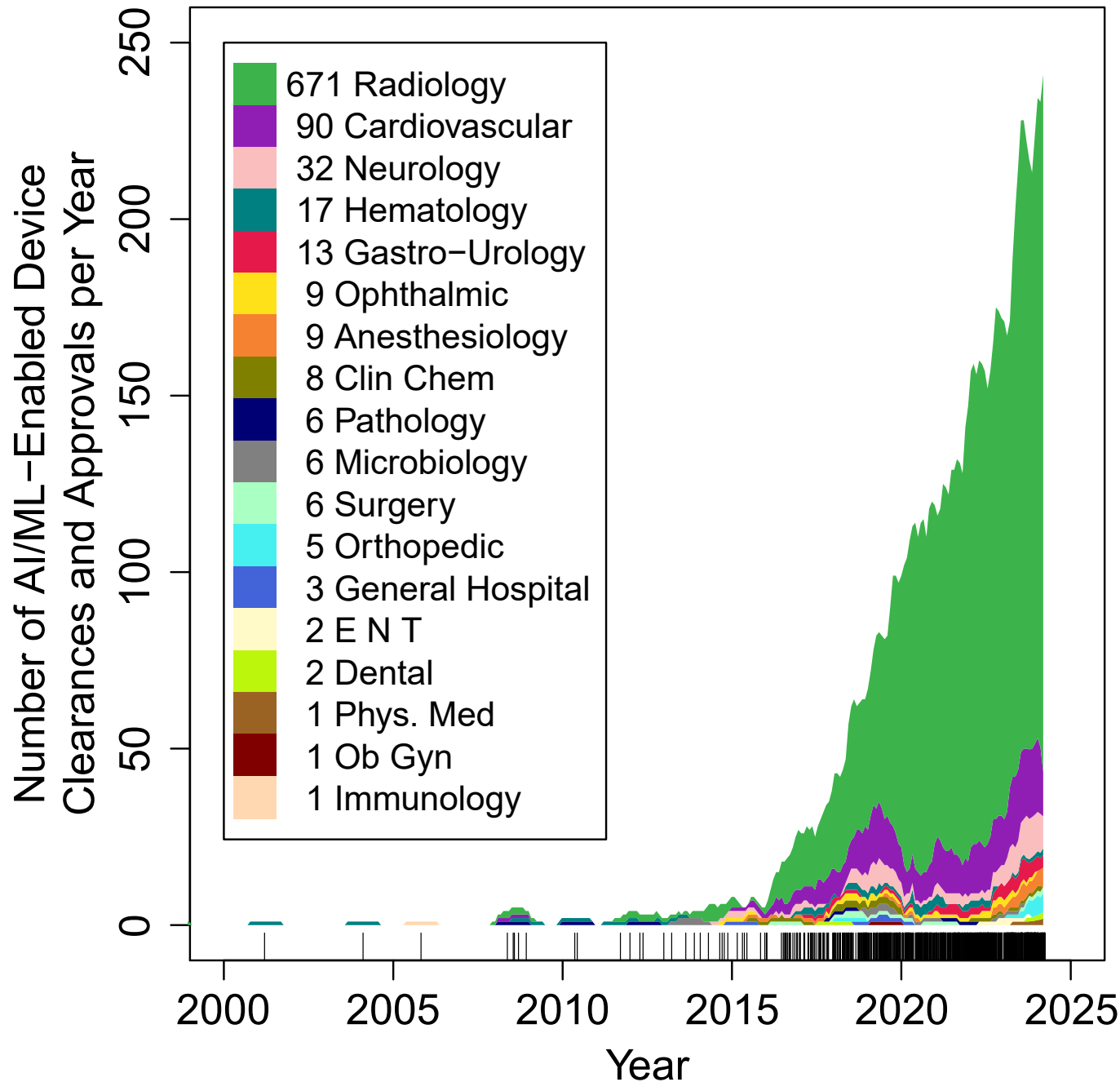
<p>VICTRE: In Silico Breast Imaging Pipeline</p> <p>Computer Model</p> <p>The Virtual Imaging Clinical Trials for Regulatory Evaluation (VICTRE) computer modeling pipeline is a set of tools that allow for the replication of...</p> <p>Medical Imaging and Diagnostics</p>	<p>Targeted Box and Blocks Test (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/>



AI/ML Devices by FDA Product Areas



Source: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-ai-ml-enabled-medical-devices>

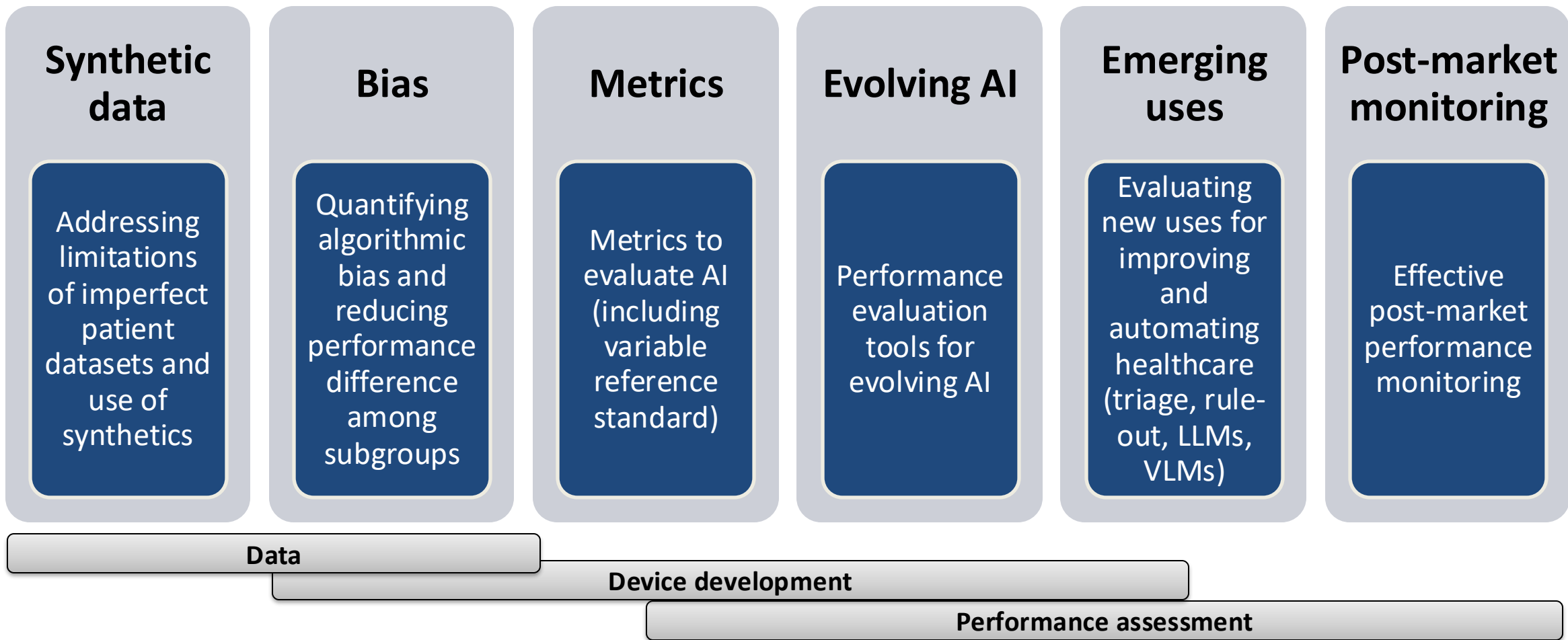
Radiology devices driven by developments in AI image processing.



OSEL/CDRH/FDA REGULATORY SCIENCE PROGRAM ON AI:

6 PROGRAM PRIORITIES

WWW.FDA.GOV/MEDICAL-DEVICES/MEDICAL-DEVICE-REGULATORY-SCIENCE-RESEARCH-PROGRAMS-CONDUCTED-OSEL/ARTIFICIAL-INTELLIGENCE-PROGRAM-RESEARCH-AIML-BASED-MEDICAL-DEVICES



It's (almost) all about the data



Medical Imaging Data Marketplace Survey Report

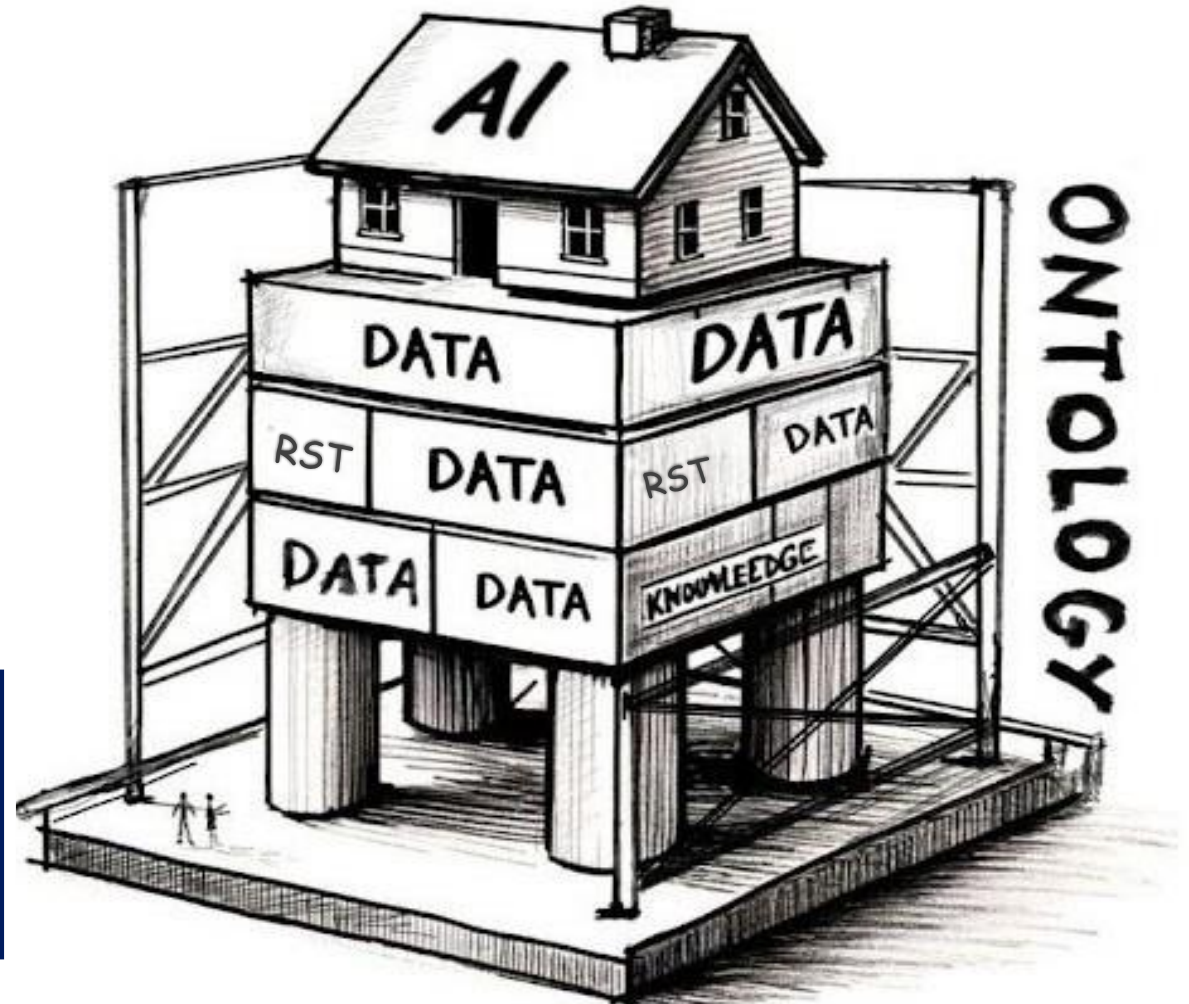
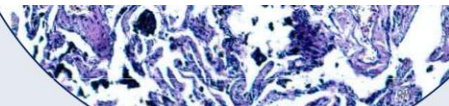
Authors: The Advanced Research Projects Agency for Health (ARPA-H)'s Investor Catalyst Hub aggregated and synthesized the results.



Data users indicated delays in generating results ranging from

2 months to 2+ years.

August 2024



Credit: Adapted from the internet

INDEX: First ARPA-H/FDA Collaborative Program



ARPA H About ▾ Research & Funding ▾ Engage & Transition ▾ News & Events ▾ Careers ▾

[Home](#) > [News & Events](#) > ARPA-H launches program to create medical imaging data exchange platform

Proposers' Day

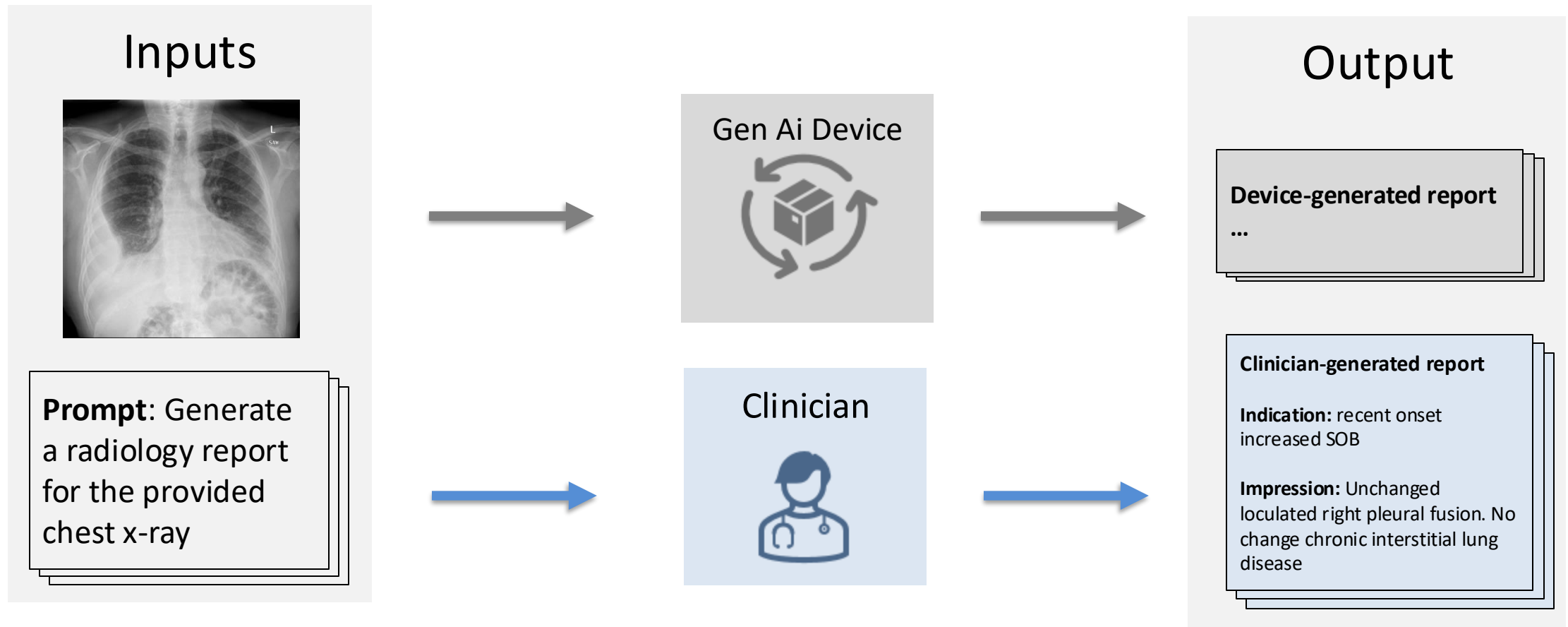
Hybrid Proposers' Day on January 9, 2025 in Seattle, WA.

ARPA-H launches program to create medical imaging data exchange platform



“Medical imaging data is scarce, expensive, siloed, and not under a practical quality system,” said INDEX Program Manager [Leana Hancu, Ph.D.](#) “INDEX is not another database; instead, it seeks to be a single, affordable, and sustainable exchange platform with built-in tools to develop regulatory-ready algorithms. The platform intends to benefit the entire imaging ecosystem, from data providers to data users and patients.”

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: **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** (with a trophy icon), **Submit** (with a rocket icon), and **Model Vote** (with a UPI icon).
- Search**: A text input field with the placeholder "Separate multiple queries with ';'".
- Select Columns to Display**: A grid of checkboxes for various metrics and attributes, including Average, IFEval, IFEval Raw, BBH, BBH Raw, MATH Lvl 5, MATH Lvl 5 Raw, GPQA, GPQA Raw, MUSR, MUSR Raw, MMLU-PRO, MMLU-PRO Raw, Type, Architecture, Precision, Not_Merged, Hub License, #Params (B), Hub (with a heart icon), Model sha, Submission Date, Upload To Hub Date, Chat Template, Generation, Base Model, and CO₂ cost (kg).
- Model types**: A list of checkboxes for model categories such as chat models (RLHF, DPO, IFT, ...), fine-tuned on domain-specific datasets, base merges and moerges, pretrained, multimodal, and continuously pretrained.
- Precision**: A list of checkboxes for precision levels: bfloat16, float16, and 4bit.
- Select the number of parameters (B)**: A range selector with values 7 and 10.
- Hide models**: A list of checkboxes for filtering models, including Deleted/incomplete, Merge/MoErge, MoE, Flagged, and Show only maintainer's highlight.



Performance Assessment Strategies: **EXPERT EVALUATION**



What is it?

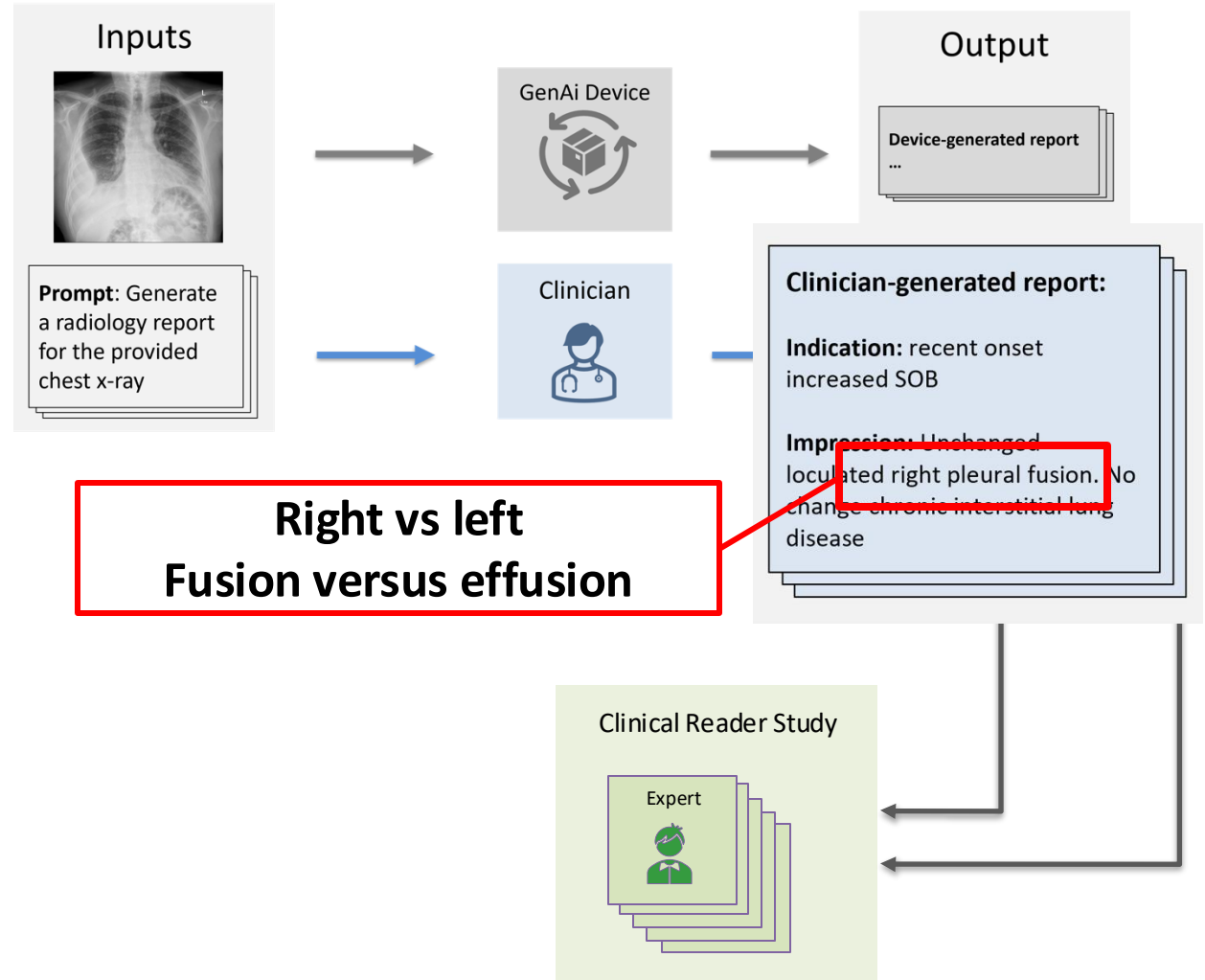
Evaluating models using expert annotations as the reference standard.

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?

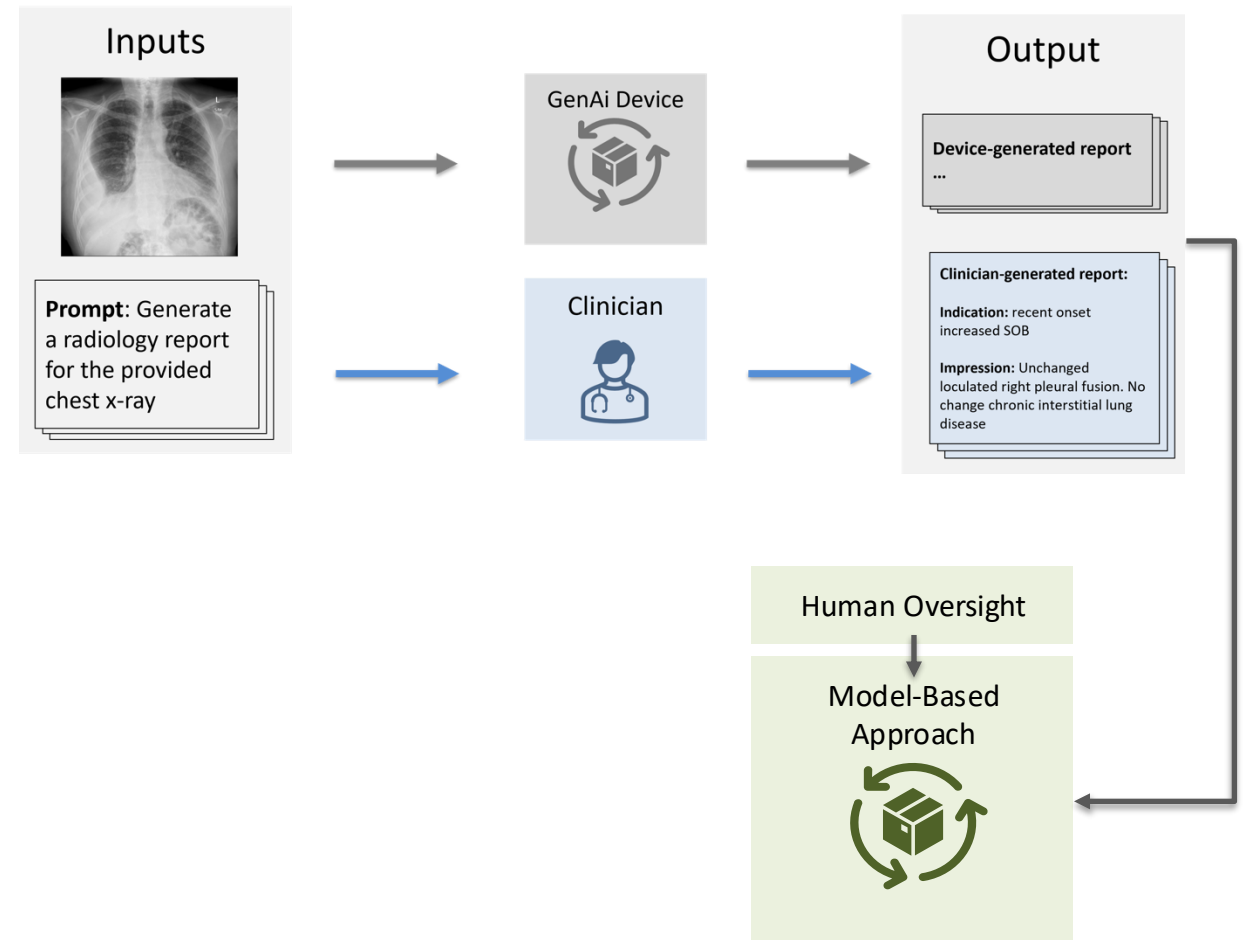
Evaluating models using a *model-based approach* (may be based on genAI) with human oversight

Advantages

- Augments human evaluation
- Scalable

Disadvantages

- Burdensome validation
- Inter-model leakage



Performance Assessment Strategies: MODEL-BASED EVALUATION



What is it?

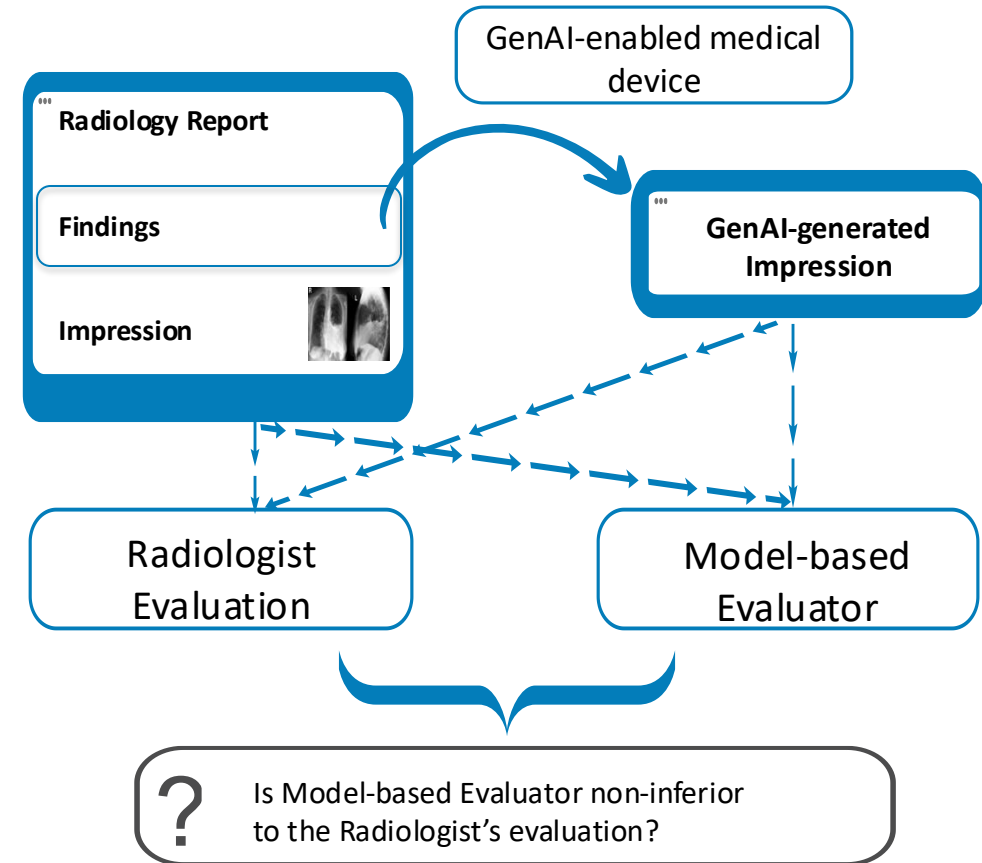
Evaluating models using a *model-based approach* (may be based on genAI) with human oversight

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: Are the findings in the GenAI-generated report found in the reference report?

Performance Assessment Strategies for genAI



Benchmarking

- 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

Expert Evaluation

- 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

Model-based evaluation

- 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

Hallucinations

2. **Confabulation:** The production of confidently stated but erroneous or false content (known colloquially as "hallucinations" or "fabrications") by which users may be misled or deceived.⁵

What are AI hallucinations?

AI hallucination is a phenomenon wherein a large language model (LLM)—often a generative AI chatbot or computer vision tool—perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate.

What are AI hallucinations?

AI hallucinations are incorrect or misleading results that AI models generate. These errors can be caused by a variety of factors, including insufficient training data, incorrect assumptions made by the model, or biases in the data used to train the model. AI hallucinations can be a problem for AI systems that are used to make important decisions, such as medical diagnoses or financial trading.

An open problem in artificial intelligence is how to train models that produce responses that are factually correct. Current language models sometimes produce false outputs or answers unsubstantiated by evidence, a problem known as "hallucinations". Language models that generate more accurate responses with fewer hallucinations are more trustworthy and can be used in a broader range of applications. To measure the factuality of language models, we are open-sourcing a new benchmark called SimpleQA.

Meaning of hallucination in English

hallucination

noun

US /həˌluː.səˈneɪʃn/ UK /həˌluː.sɪˈneɪʃn/

hallucination noun (HUMANS)

[C or U]

the experience of seeing, hearing, feeling, or smelling something that does not exist, usually because of a health condition or because you have taken a drug:

- A high temperature can cause hallucinations.
- Auditory hallucination is more common than people think.
- He was suffering from drug-induced hallucination.

[C]

something that you see, hear, feel or smell that does not exist:

- She had been alone for so long that when she saw the boat coming to rescue her she was convinced it was a hallucination.

— SMART Vocabulary: related words and phrases

Dreaming

be hearing/imagining/seeing things idiom	be miles away idiom
daydream	daydreamer
daydreaming	dream
dreamfully	dreamless
dreamlessly	dreamscape
fantasist	imagination
in a dream idiom	mile
night terror	nightmare
oneiric	redream
reverie	Walter Mitty

[See more results »](#)

hallucination noun (COMPUTERS)

[C]

false information that is produced by an artificial intelligence (= a computer system that has some of the qualities that the human brain has, such as the ability to produce language in a way that seems human):

- If the chatbot is used in the classroom as a teaching aid, there is a risk that its hallucinations will enter the permanent record.
- Because large language models are designed to produce coherent text, their hallucinations often appear plausible.
- She discovered that the articles cited in the essay did not exist, but were hallucinations that had been invented by the AI.

[U]

the fact of an artificial intelligence (= a computer system that has some of the qualities that the human brain has, such as the ability to produce language in a way that seems human) producing false information:

- The system tends to make up information when it doesn't know the exact answer – an issue known as hallucination.
- Is it possible to solve the problem of AI hallucination?



hallucination noun

hal·lu·ci·na·tion (həˌlʊːsəˈnāːʃən)

plural hallucinations

Synonyms of hallucination >

- a** : a sensory perception (such as a visual image or a sound) that occurs in the absence of an actual external stimulus and usually arises from neurological disturbance (such as that associated with delirium tremens, schizophrenia, Parkinson's disease, or narcolepsy) or in response to drugs (such as LSD or phencyclidine)
visual/auditory/olfactory/gustatory/tactile hallucinations
a drug-induced hallucination
An important aspect of the study of hallucinations is the judgement of reality. How does a patient confer the character of reality on stimuli which, beyond any reasonable doubt, originate in his own mind?
– Cesare Davalli et al.
- b** : the object of a hallucinatory perception
wasn't sure if the creature was real or a hallucination
- 2** : an unfounded or mistaken impression or notion : **DELUSION**
... that popular hallucination, from which not even great scientists are ... free ...
– Lewis Mumford
- 3** **computing** : a plausible but false or misleading response generated by an artificial intelligence algorithm
"This type of artificial intelligence we're talking about can sometimes lead to something we call hallucination," said Prabhakar Raghavan in an interview with Germany's Welt am Sonntag newspaper published on Saturday. "This is then expressed in such a way that a machine delivers a convincing but completely fictitious answer."
– Matthew Broersma

Generative AI suffers from fakes (hallucinations)



DOWNLOAD PDF

92 12

Fake detection in AI-assisted image recovery using scanning Fourier Ring Correlation (sFRC)

SIGNAL PROCESSING AND ANALYSIS

ARTIFACTS BIOMEDICAL IMAGING DEEP LEARNING FAKE DETECTION HALLUCINATION

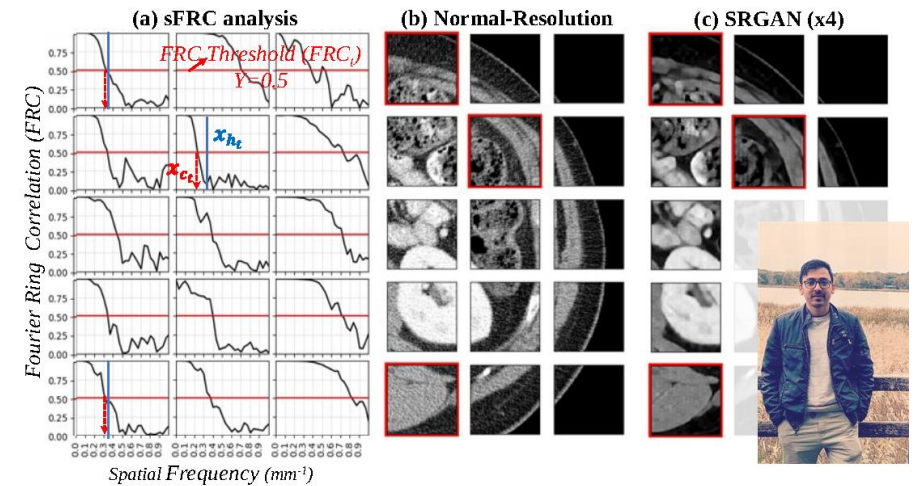
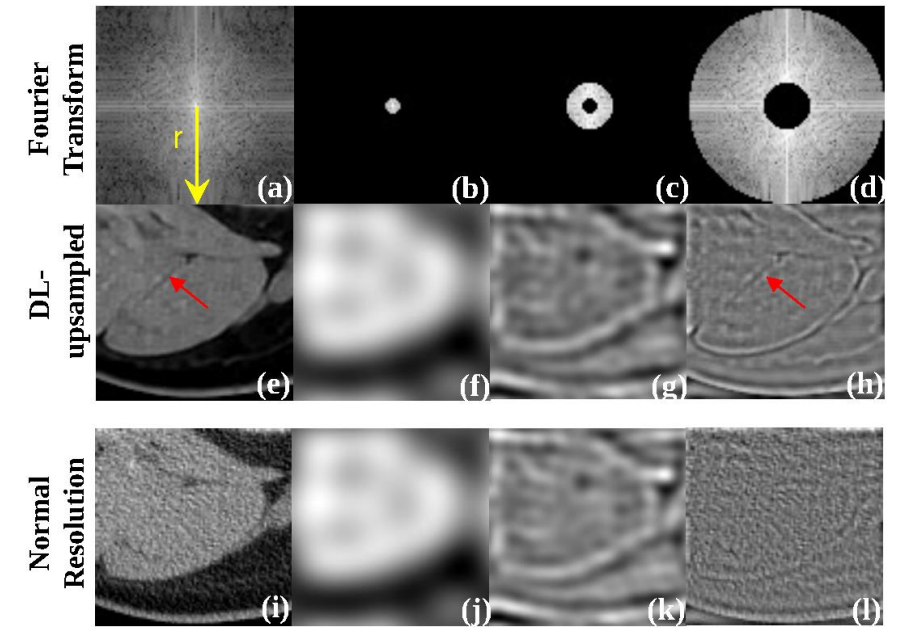
IMAGE QUALITY INVERSE PROBLEM MACHINE LEARNING SUBSAMPLED ACQUISITION

SUPER-RESOLUTION

Prabhat Kc, Rongping Zeng, Nirmal Soni, aldo badano

Abstract

Deep learning (DL) methods are currently being explored to recover images from sparse-view, limited-data, and undersampled acquisitions in medical applications. Although DLbased solutions may appear visually appealing based on likability/subjective criteria (such as less noise, smooth features), they may also suffer from imperceptible fakes. This issue is further exacerbated by a lack of easy-to-use techniques and robust metrics for the identification of fakes in DL-based outputs. In this work, we propose performing Fourier Ring Correlation (FRC)based analysis over small patches and concomitantly scanning across DL-based outputs and their reference counterparts to identify fakes. We term the metrics as sFRC. We describe the rationale behind sFRC and provide its mathematical framework. The thresholds required for the sFRC can be set using predefined fake features or imaging theory-based fake maps. We use sFRC to identify fakes for two undersampled medical imaging problems (CT super-resolution and MRI subsampled recovery). We demonstrate the effectiveness of sFRC in finding fake features for the two imaging problems and its agreement with a different imaging theory-based method on fake feature maps. Finally, we quantify the incidences of fakes from DL-based methods relative to indistribution versus out-of-distribution data and the increment in subsampling rate.



Summary of Regulatory Science Challenges in Medical AI

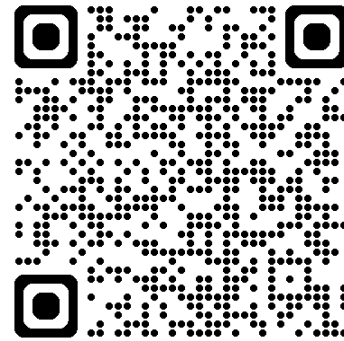


Accessible and sustainable data platforms aligned with regulatory requirements

Evaluation platforms including new evaluation methodologies and new performance metrics

Synthetic data: How to evaluate quality of synthetic datasets?

Thank you for your attention



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



**U.S. FOOD & DRUG
ADMINISTRATION**