U.S. FOOD & DRUG ADMINISTRATION

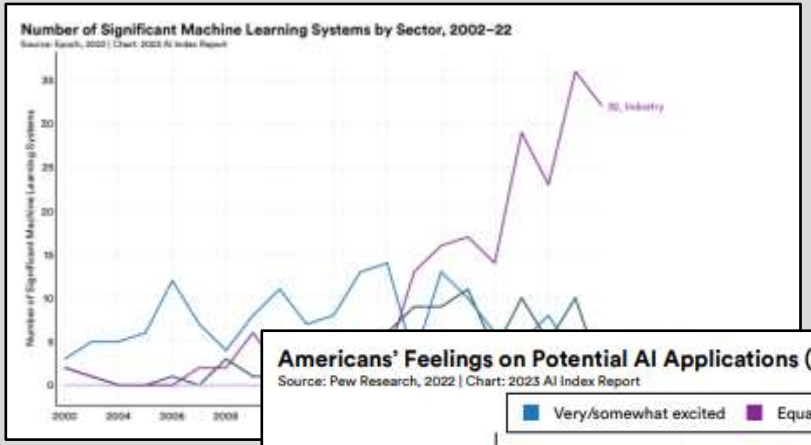
Addressing Regulatory Science Gaps in Artificial Intelligence (AI) / Machine Learning (ML)

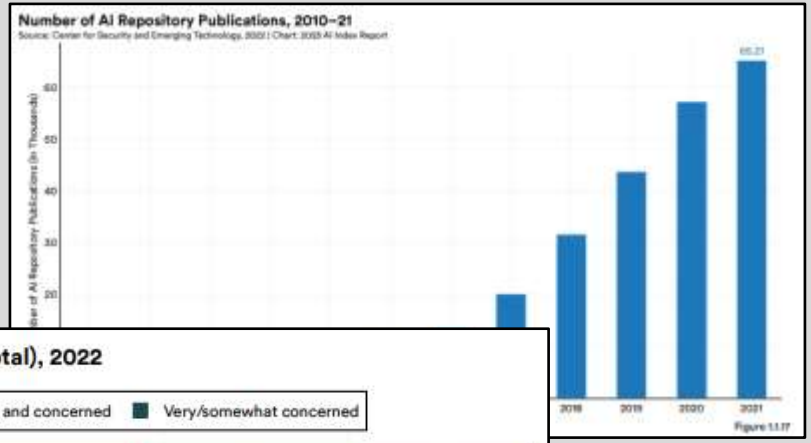
FDA Small Business Regulatory Education for Industry (REdI) Annual Conference

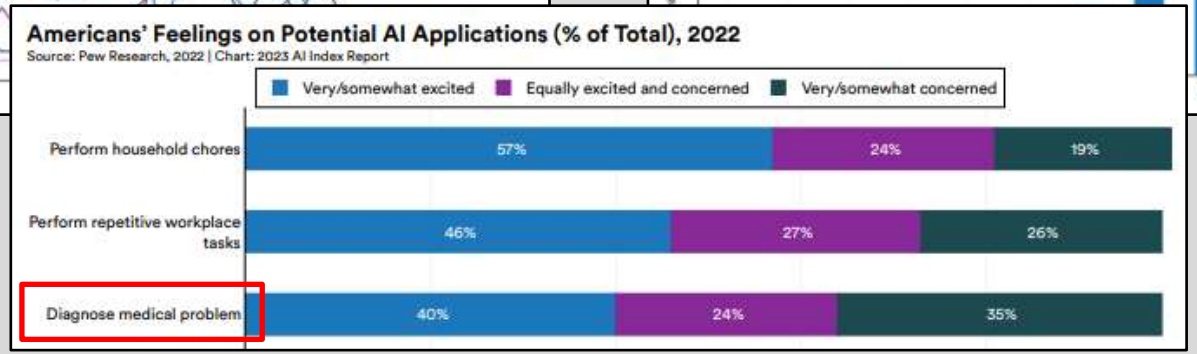
June 8, 2023

Alexej Gossmann
Staff Fellow
Division of Imaging, Diagnostics and Software Reliability
Office of Science and Engineering Labs
Center for Devices and Radiological Health
U.S. Food and Drug Administration

AI/ML is taking the world by storm!

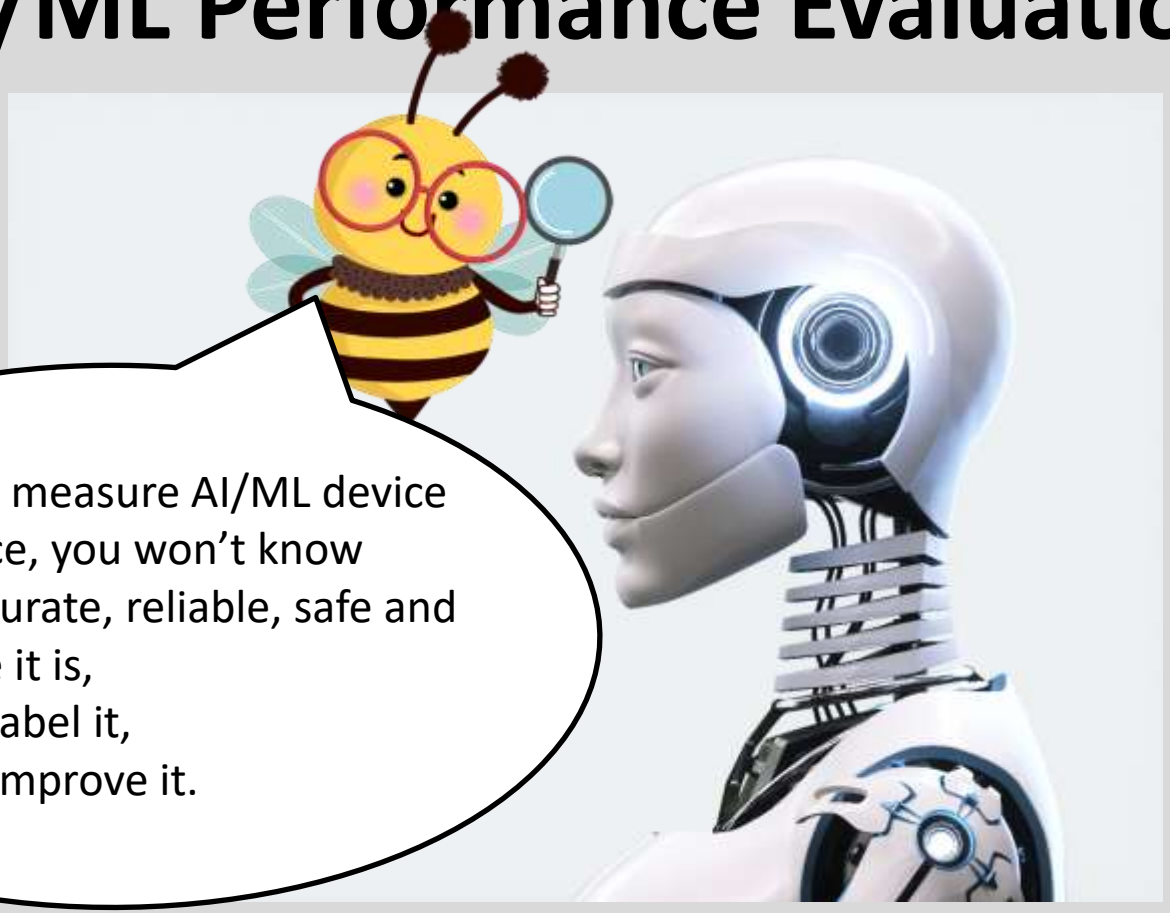






Stanford HAI. 2023 AI Index Report: Measuring Trends in Artificial intelligence. Retrieved 4 April 2023.
aiindex.stanford.edu/report/

AI/ML Performance Evaluation



If you don’t measure AI/ML device performance, you won’t know

- How accurate, reliable, safe and effective it is,
- How to label it,
- How to improve it.

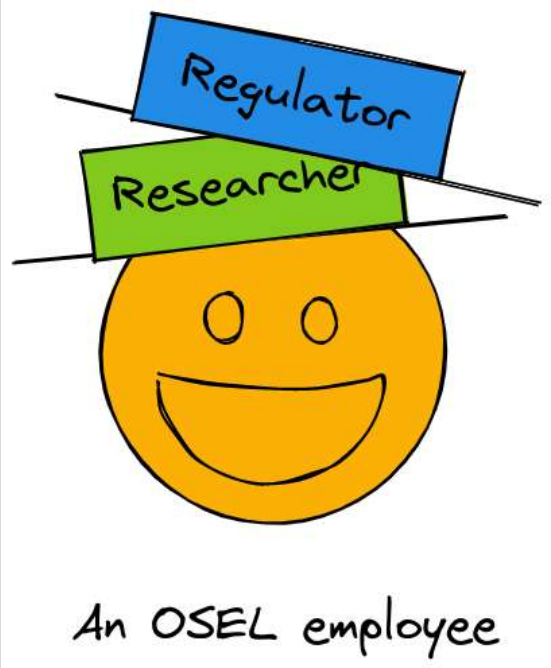
Learning Objectives

- Explain who we are at OSEL, DIDSR
- Describe regulatory science challenges and gaps in medical AI/ML
- Describe OSEL AI/ML research program activities to address these gaps

Office of Science and Engineering Labs (OSEL)

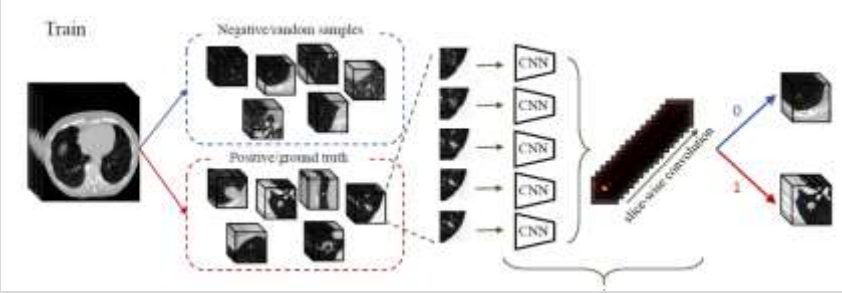
Mission Statement

Accelerating patient access to innovation, safe and effective medical devices through best-in-the-world regulatory science.



What OSEL/DIDSR Does

- Division of Imaging, Diagnostics, and Software Reliability (DIDSR)
- Conduct regulatory science research for a variety of imaging, AI/ML, MXR, and diagnostic devices.
- Develop approaches for assessing imaging and big-data technologies.



A Collaborative Approach to AI/ML-enabled devices at CDRH

Recent Milestones

Published AI/ML-SaMD Discussion Paper

First joined Collaborative Community related to AI/ML

Evolving Role of AI in Radiological Imaging

Public Workshop on AI/ML in Radiological Imaging

PEAC Mtg on Patient Trust in AI/ML Devices

Published AI/ML Medical Device Software Action Plan

Posted List of Currently Marketed AI/ML Devices

Transparency of AI/ML-Enabled Medical Devices

Public Workshop on Transparency of AI/ML Devices

Published GMLP Guiding Principles

Published PCCP for AI/ML Draft Guidance

2019 **2020** **2021** **2023**

A Collaborative Approach to AI/ML-enabled devices at CDRH

Recent Milestones

Published AI/ML-SaMD Discussion Paper

First joined Collaborative Community related to AI/ML

Evolving Role of AI in Radiological Imaging

Public Workshop on AI/ML in Radiological Imaging

PEAC Mtg on Patient Trust in AI/ML Devices

2019

2020

Published AI/ML Medical Device Software Action Plan

Posted List of Currently Marketed AI/ML Devices

Transparency of AI/ML-Enabled Medical Devices

Public Workshop on Transparency of AI/ML Devices

Published GMLP Guiding Principles

Published PCCP for AI/ML Draft Guidance

2021

2023

2021+: AI/ML Medical Device Software Action Plan *www.fda.gov/media/145022/download*

- ❑ Update the proposed AI/ML framework
- ❑ Strengthen FDA's role in harmonizing GMLP
- ❑ Foster a patient-centered approach
- ❑ Support development of regulatory science methods
- ❑ Advance real-world performance pilots

Knowledge Check

What are some reasons to measure the performance of AI/ML-enabled medical devices?

1. Ensure that these systems are safe and effective
2. Characterize accuracy and precision, across a diverse patient population
3. Labeling
4. Understand how to improve the AI device
5. All of the above

Knowledge Check

What are some reasons to measure the performance of AI/ML-enabled medical devices?

1. Ensure that these systems are safe and effective
2. Characterize accuracy and precision, across a diverse patient population
3. Labeling
4. Understand how to improve the AI device
5. All of the above

OSEL AI/ML Program

- **AI/ML program**
  - Regulatory science research
  - Developing robust AI/ML test methods
  - Evaluating methodologies for assessing AI/ML

- **AI/ML team identified regulatory gaps**
  - Not all AI/ML knowledge gaps
  - Important ones to support FDA regulatory mission

Regulatory Science Gaps and Challenges

- Limited labeled training and test data
- Bias, equity, and generalizability
- Ground truth and metrics for performance estimation
- Evolving algorithms – How to maintain safety and effectiveness for devices with a predetermined change control plan (PCCP)
- Emerging clinical application of AI/ML
- Data Drift and Postmarket AI/ML Performance Monitoring

Regulatory Science Gaps and Challenges

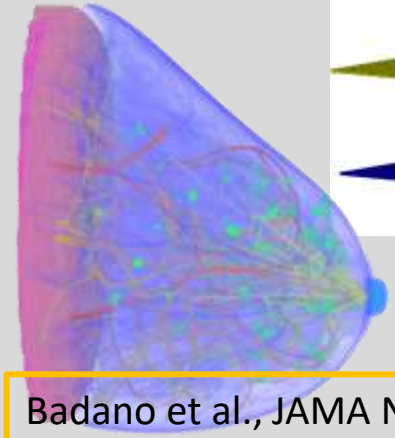
- **Limited labeled training and test data**
- Bias, equity, and generalizability
- Ground truth and metrics for performance estimation
- Evolving algorithms – How to maintain safety and effectiveness for devices with a predetermined change control plan (PCCP)
- Emerging clinical application of AI/ML
- Data Drift and Postmarket AI/ML Performance Monitoring

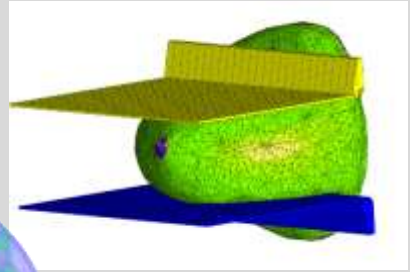
Limited labeled training and test data

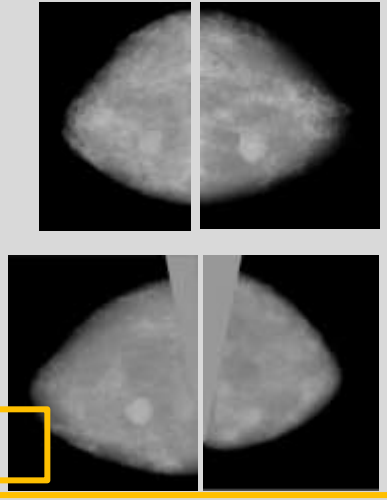
- **Need for:**
  - Fundamental understanding of limitations of smaller datasets; and
  - Novel techniques to enhance AI/ML algorithm training and testing when real-world datasets are limited in size

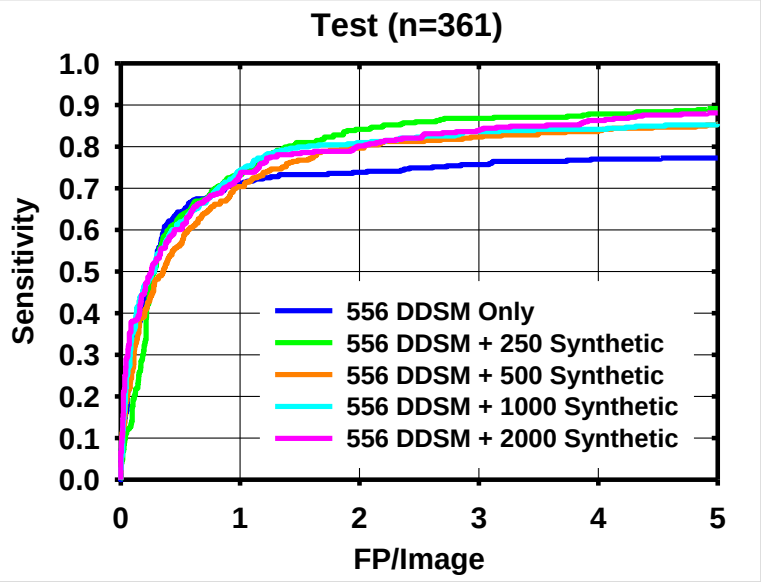
Use of synthetic data for AI training and testing

- AI algorithms require large training data sets for high performance
- Limited annotated data sets for medical images
- In-silico images may help









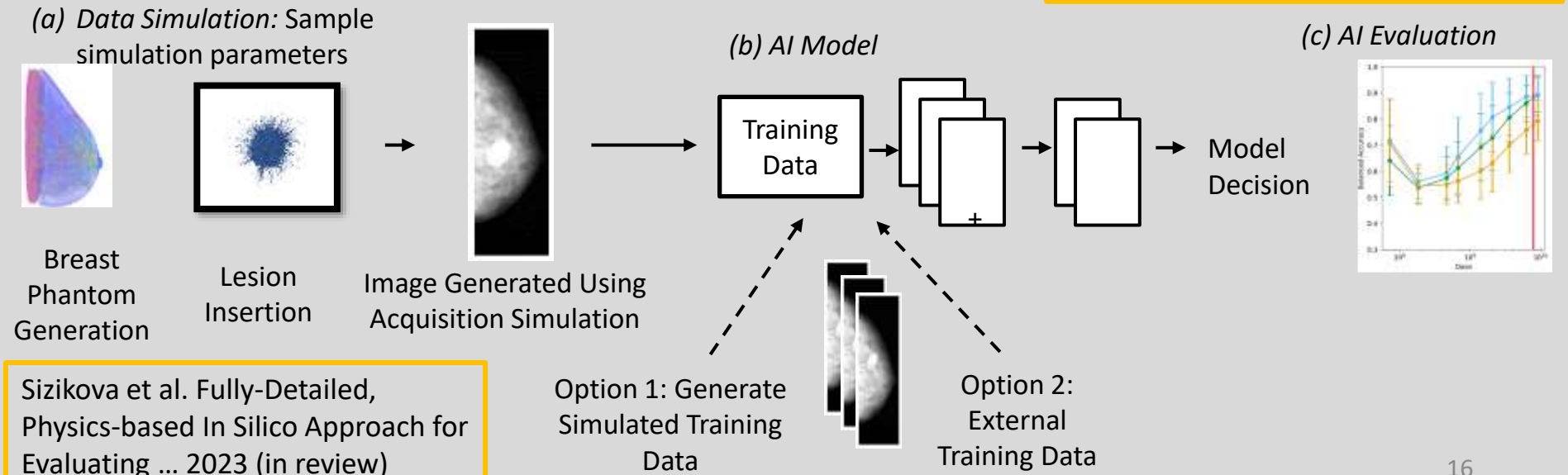
Badano et al., JAMA Network Open 2018

Cha et al., “Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning,” Journal of Medical Imaging 2020

REALYSM: Simulations-based testing for AI devices

Goal: Generate realistic simulated data where real patient examples are unavailable

Badano et al. The stochastic digital human ... ArXiv preprint 2023.



Sizikova et al. Fully-Detailed, Physics-based In Silico Approach for Evaluating … 2023 (in review)

Regulatory Science Gaps and Challenges

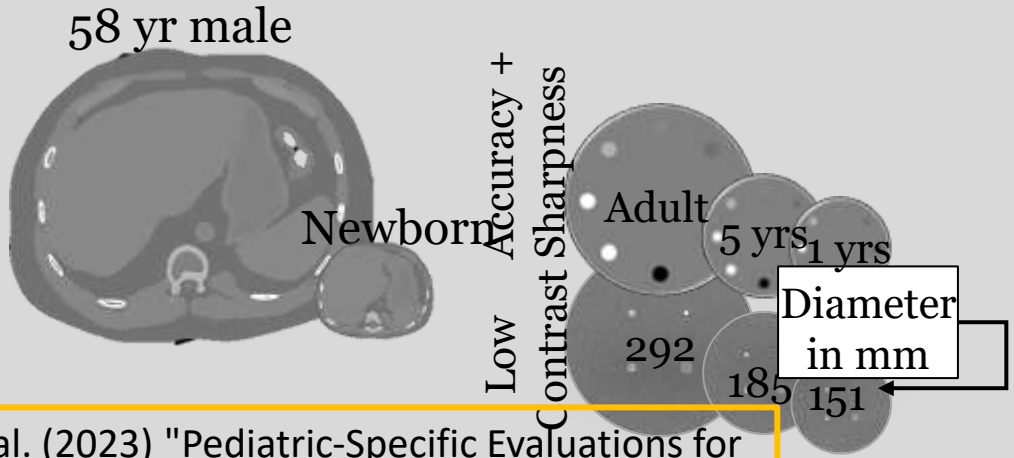
- Limited labeled training and test data
- **Bias, equity, and generalizability**
- Ground truth and metrics for performance estimation
- Evolving algorithms – How to maintain safety and effectiveness for devices with a predetermined change control plan (PCCP)
- Emerging clinical application of AI/ML
- Data Drift and Postmarket AI/ML Performance Monitoring

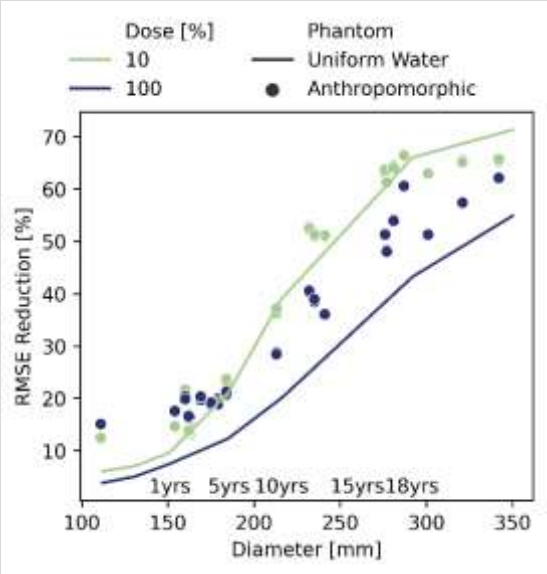
Bias, equity, and generalizability

- There is a need for methods to understand, analyze and minimize performance differences of AI/ML-enabled devices among subgroups

Pediatric-Specific Evaluations for Deep Learning CT Image Reconstruction and Denoising

- Deep learning image reconstruction (DLIR) models primarily trained on adults.
- Do pediatric patients benefit equally from adult-trained DLIR models?
- **PEDiatric CT Evaluation ToolKit (PED-ETK)**





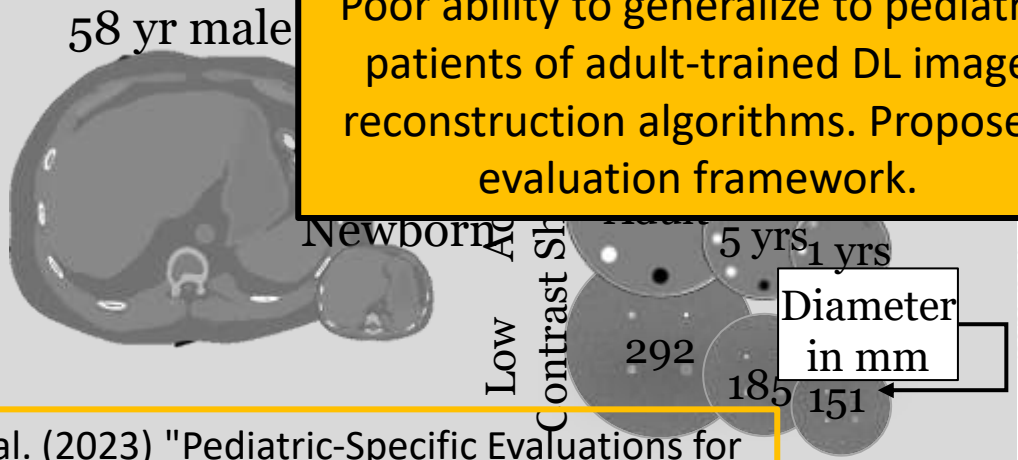
Nelson et al. (2023) "Pediatric-Specific Evaluations for Deep Learning CT Image Reconstruction and Denoising Techniques" - under review

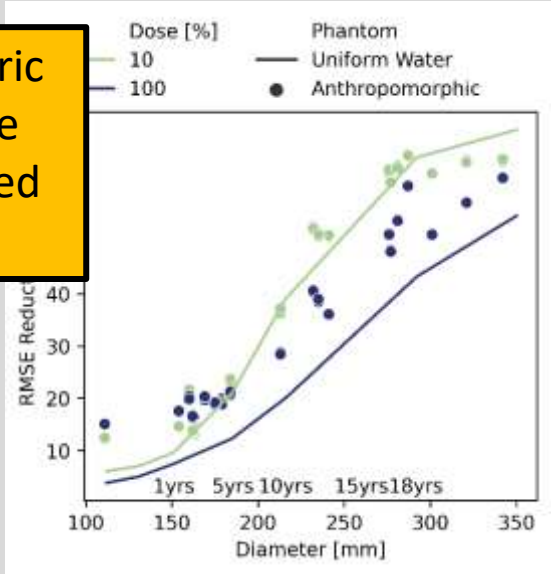
github.com/DIDSR/PED-ETK

Pediatric-Specific Evaluations for Deep Learning CT Image Reconstruction and Denoising

- Deep learning image reconstruction (DLIR) models primarily trained on adults.
- Do pediatric patients benefit equally from adult-trained DLIR models?
- **PEDiatric CT Evaluation ToolKit (PED-ETK)**

Poor ability to generalize to pediatric patients of adult-trained DL image reconstruction algorithms. Proposed evaluation framework.





Nelson et al. (2023) "Pediatric-Specific Evaluations for Deep Learning CT Image Reconstruction and Denoising Techniques" - under review

github.com/DIDSR/PED-ETK

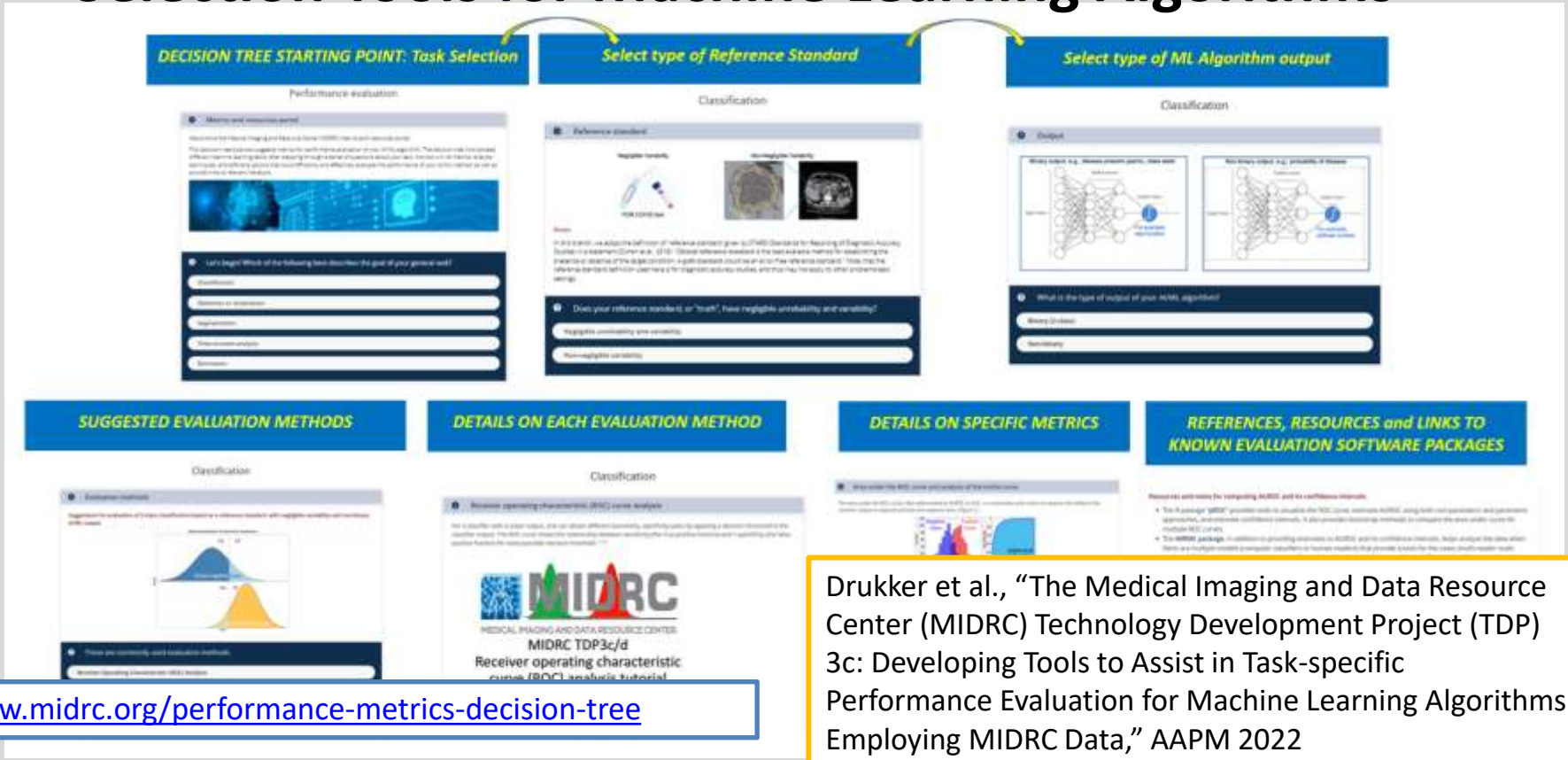
Regulatory Science Gaps and Challenges

- Limited labeled training and test data
- Bias, equity, and generalizability
- **Ground truth and metrics for performance estimation**
- Evolving algorithms – How to maintain safety and effectiveness for devices with a predetermined change control plan (PCCP)
- Emerging clinical application of AI/ML
- Data Drift and Postmarket AI/ML Performance Monitoring

Ground truth and metrics for performance estimation

- A need to understand how to determine the level of truth needed to evaluate AI-enabled devices in a least burdensome fashion
- Metrics used to determine AI/ML performance
- Determination of acceptable performance criteria

MIDRC: Task-specific Performance Evaluation Metric Selection Tools for Machine Learning Algorithms



www.midrc.org/performance-metrics-decision-tree

Drukker et al., "The Medical Imaging and Data Resource Center (MIDRC) Technology Development Project (TDP) 3c: Developing Tools to Assist in Task-specific Performance Evaluation for Machine Learning Algorithms Employing MIDRC Data," AAPM 2022

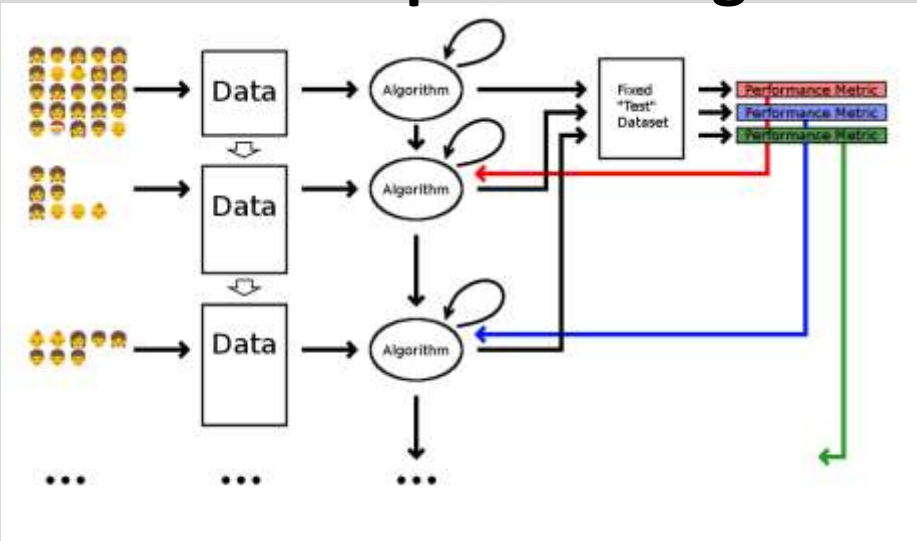
Regulatory Science Gaps and Challenges

- Limited labeled training and test data
- Bias, equity, and generalizability
- Ground truth and metrics for performance estimation
- **Evolving algorithms – How to maintain safety and effectiveness for devices with a predetermined change control plan (PCCP)**
- Emerging clinical application of AI/ML
- Data Drift and Postmarket AI/ML Performance Monitoring

Evolving algorithms

- How to maintain safety and effectiveness for devices with a predetermined change control plan (PCCP)
- Our stakeholders would like a more flexible pre-market regulatory process to allow for periodic modifications of AI/ML algorithms over time and evolving AI algorithms without the need for a new regulatory submission.
- Many open questions related to the regulation of such devices.

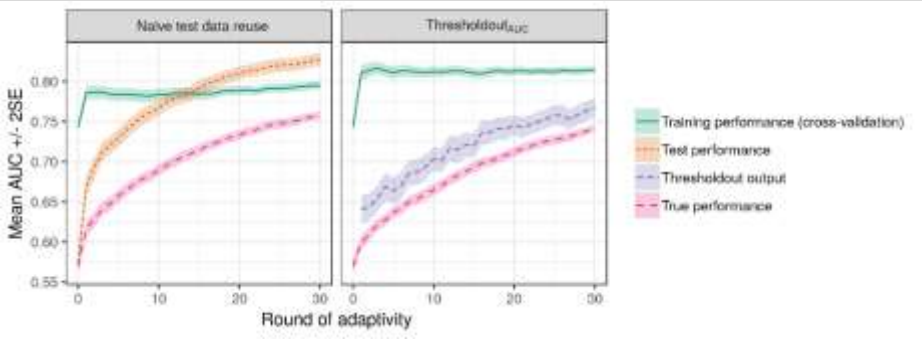
How can we reuse an existing test dataset to validate sequential algorithmic modifications?



Methods that allow for valid test data reuse restrict the amount of information leaked with each query by

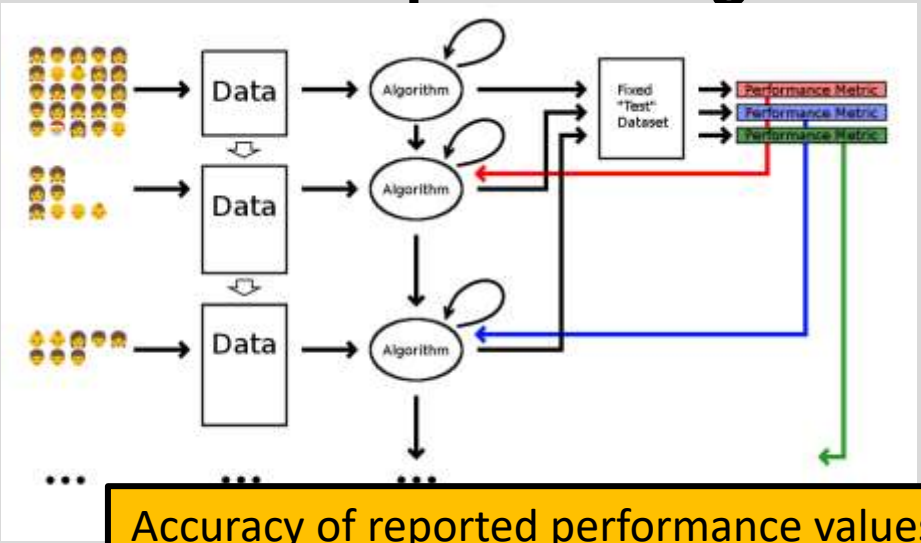
(a) perturbing the query result with random noise → differential privacy

(b) restricting the number of bits of information returned.



Gossmann et al., "Test Data Reuse for the Evaluation …," SIAM J Math Data Science, 2021

How can we reuse an existing test dataset to validate sequential algorithmic modifications?

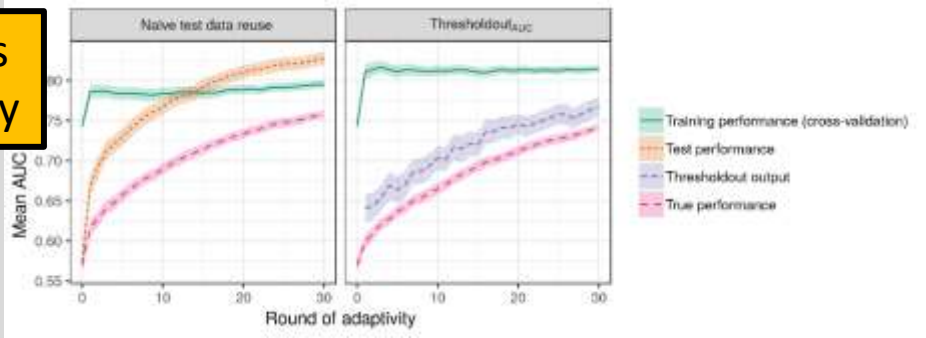


Methods that allow for valid test data reuse restrict the amount of information leaked with each query by

(a) perturbing the query result with random noise → differential privacy

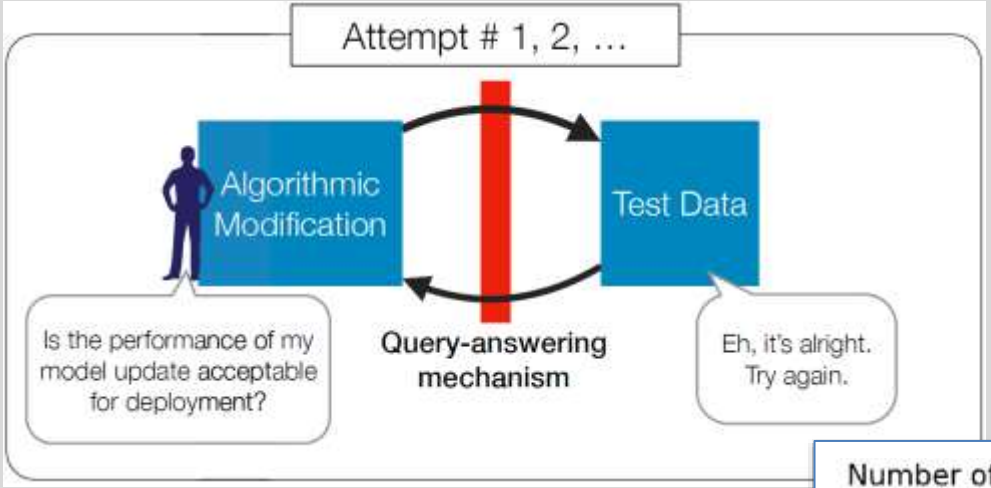
(b) restricting the number of bits of information returned.

Accuracy of reported performance values improves at the cost of higher uncertainty



Gossmann et al., "Test Data Reuse for the Evaluation …," SIAM J Math Data Science, 2021

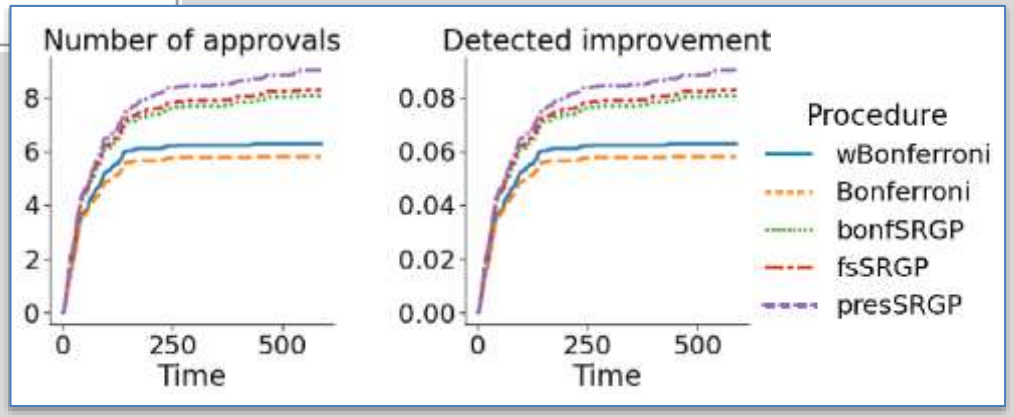
How can we reuse an existing test dataset to validate sequential algorithmic modifications?



Methods that allow for valid test data reuse restrict the amount of information leaked with each query by

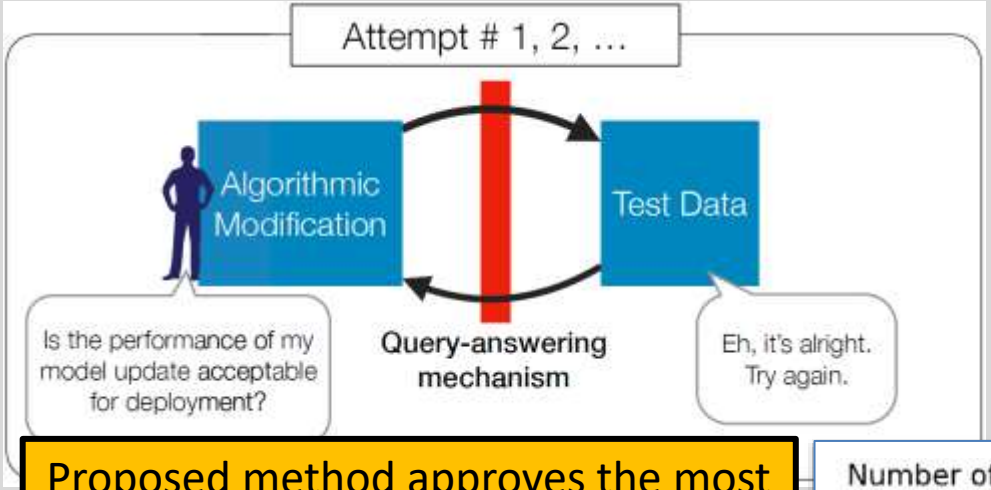
(a) perturbing the query result with random noise → differential privacy

(b) restricting the number of bits of information returned.



Feng et al., "Sequential algorithmic modification with test data reuse," UAI, 2022

How can we reuse an existing test dataset to validate sequential algorithmic modifications?

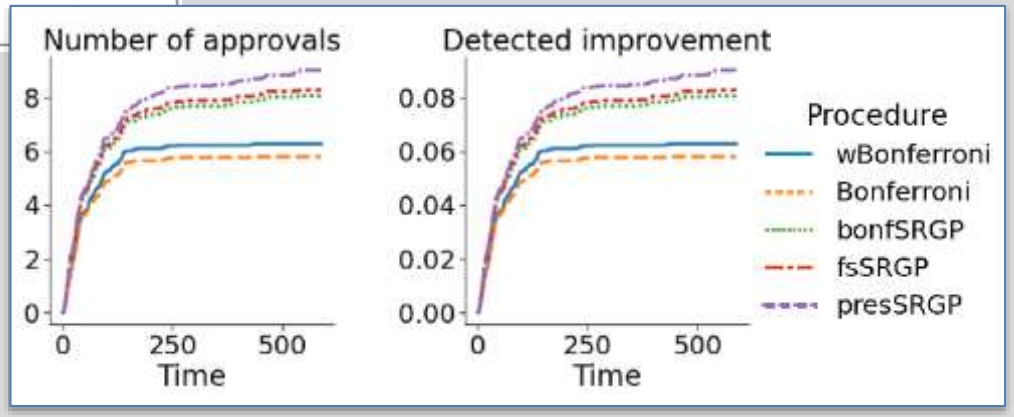


Methods that allow for valid test data reuse restrict the amount of information leaked with each query by

(a) perturbing the query result with random noise → differential privacy

(b) restricting the number of bits of information returned.

Proposed method approves the most model updates and achieves the best performance, while controlling the rate of bad approvals.



Feng et al., "Sequential algorithmic modification with test data reuse," UAI, 2022

Regulatory Science Gaps and Challenges

- Limited labeled training and test data
- Bias, equity, and generalizability
- Ground truth and metrics for performance estimation
- Evolving algorithms – How to maintain safety and effectiveness for devices with a predetermined change control plan (PCCP)
- **Emerging clinical application of AI/ML**
- Data Drift and Postmarket AI/ML Performance Monitoring

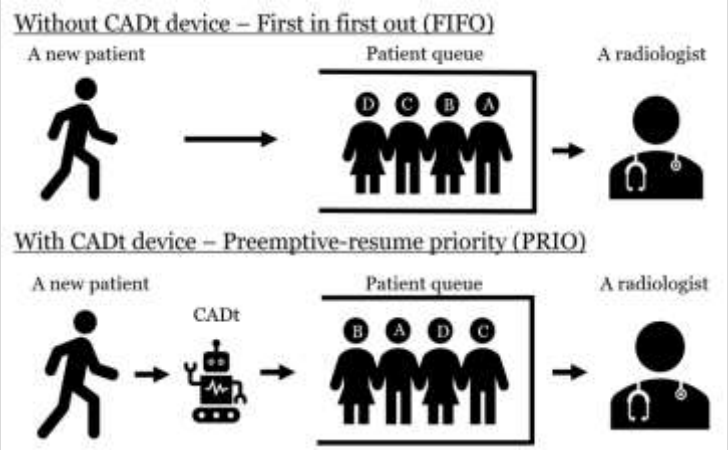
Emerging clinical application of AI/ML

- Device sponsors continue to think of new ways to utilize AI/ML in medical practice, including:
  - Automating patient referrals,
  - Triaging patients,
  - Reading images autonomously,
  - Large language models (LLMs) applied to medical records,
  - Etc.
- We need methods for evaluating these new and different uses of AI

A Modeling Tool for Streamlined Assessment of Emerging Radiological Computer-Assisted Triage (CADt) and Notification Software

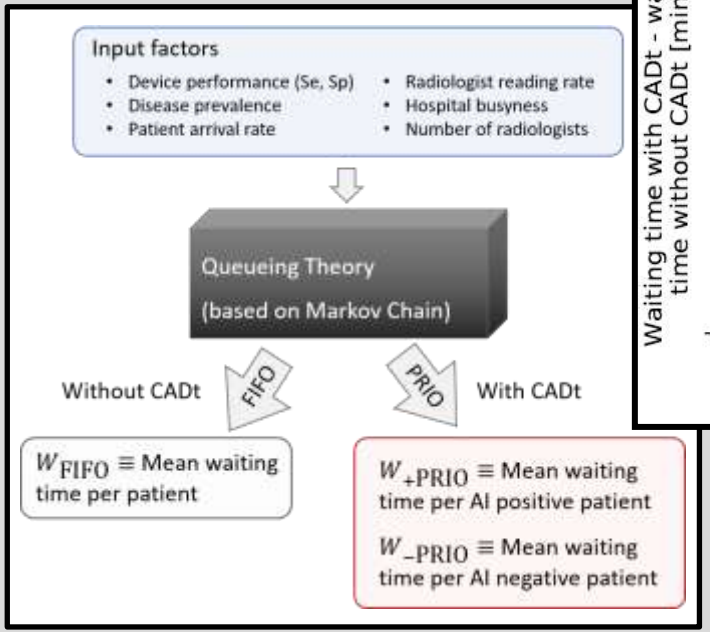
- 30+ FDA-approved CADt devices since 2018
- Why CADt devices?
  - Faster diagnosis and treatment for time sensitive diseases e.g. stroke
- How effective is a CADt device?

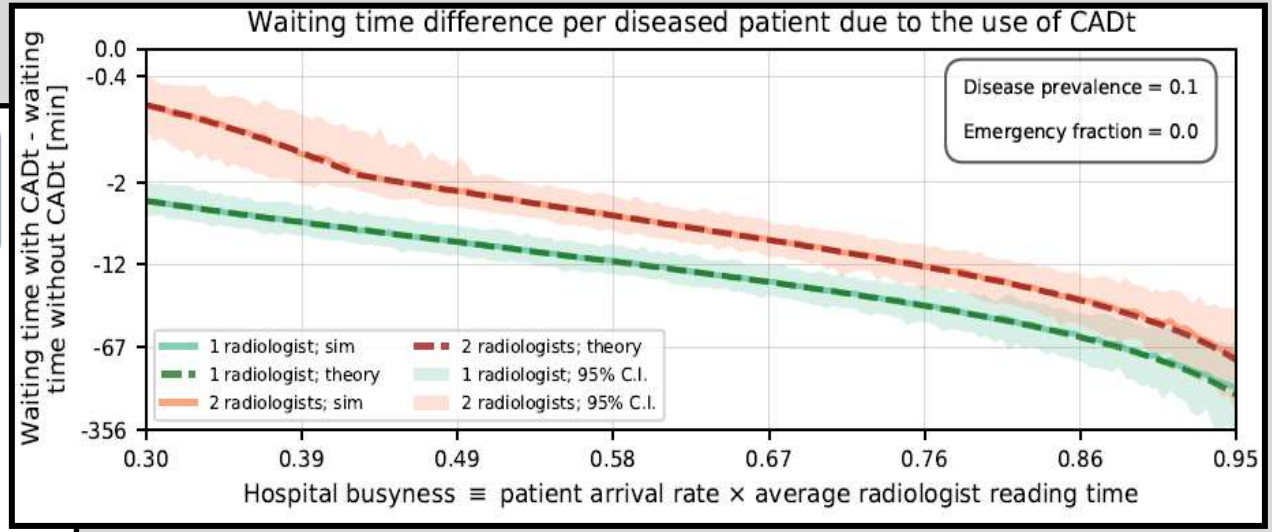
  → Use queueing theory to quantify the amount of time savings



Thompson et al., "Wait-Time-Saving Analysis and …," SPIE MI, 2022

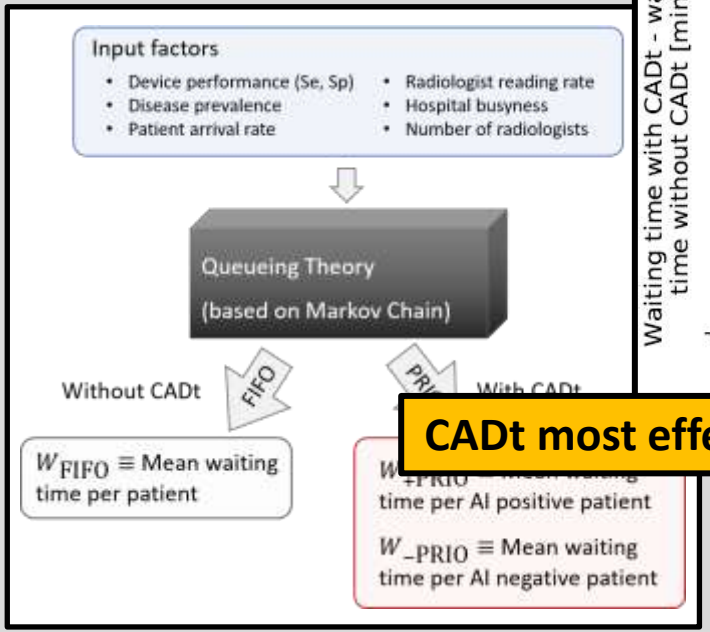
A Modeling Tool for Streamlined Assessment of Emerging Radiological Computer-Assisted Triage (CADt) and Notification Software

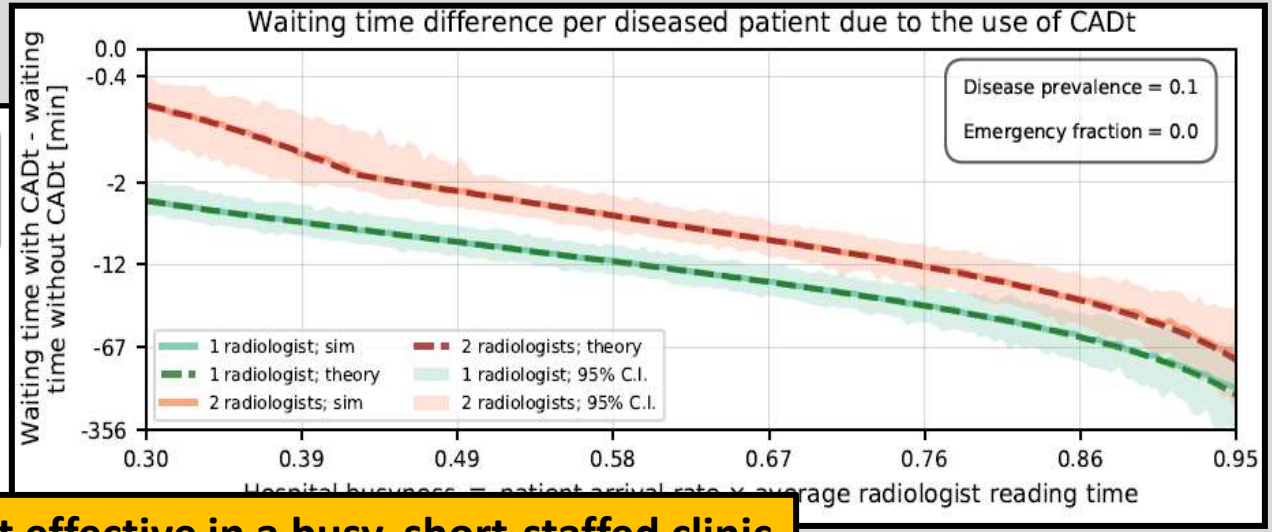




Thompson et al., "Wait-Time-Saving Analysis and …," SPIE MI, 2022

A Modeling Tool for Streamlined Assessment of Emerging Radiological Computer-Assisted Triage (CADt) and Notification Software





CADt most effective in a busy, short-staffed clinic

Thompson et al., "Wait-Time-Saving Analysis and …," SPIE MI, 2022

Regulatory Science Gaps and Challenges

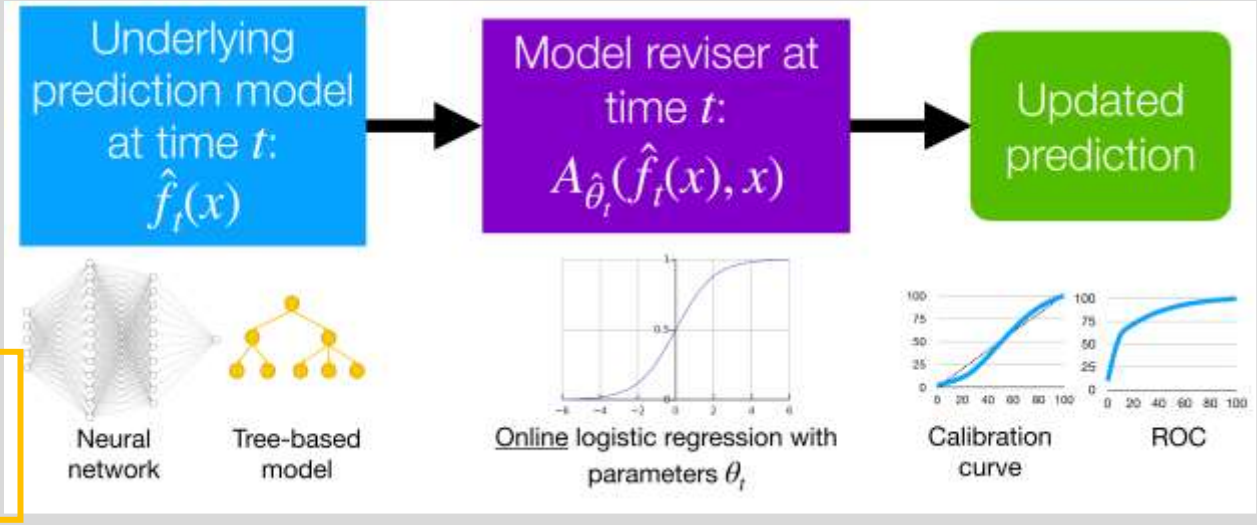
- Limited labeled training and test data
- Bias, equity, and generalizability
- Ground truth and metrics for performance estimation
- Evolving algorithms – How to maintain safety and effectiveness for devices with a predetermined change control plan (PCCP)
- Emerging clinical application of AI/ML
- **Data Drift and Postmarket AI/ML Performance Monitoring**

Data Drift and Postmarket AI/ML Performance Monitoring

- Data acquisition systems and protocols, and patient populations change over time and by site
- AI/ML device users, such as radiologists, and patients want to know that the AI products they are using will be accurate and reliable even as practice and patient populations change
- We need planned and standardized methods for detecting changes to the inputs of AI devices, monitoring the accuracy of their outputs, and mitigating effects of those drifts

Online Recalibration

- Model updates can protect against changes in the environment, and learn from accumulating data.
- However, algorithmic modifications also carry the risk of deteriorating model performance.
- We design an online logistic recalibration and revision procedure that provides performance guarantees.



Feng et al., "Bayesian logistic regression for online recalibration and revision …," JAMIA 2022.

Knowledge Check

Which of the following can be considered emerging applications of AI/ML in medical imaging?

1. Computer aided diagnosis or detection AI systems
2. Autonomous AI systems for patient referral
3. Systems for patient triage, rule-in, or rule-out

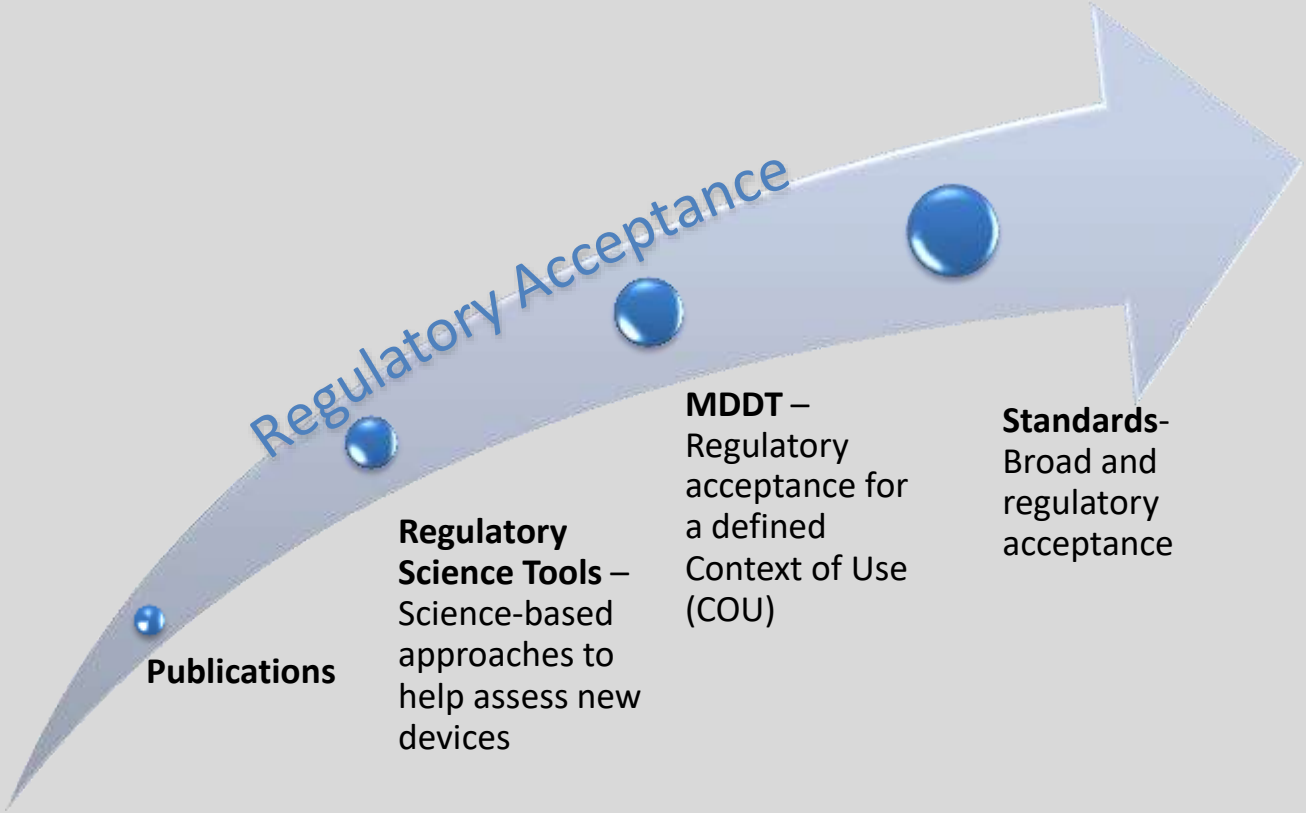
Knowledge Check

Which of the following can be considered emerging applications of AI/ML in medical imaging?

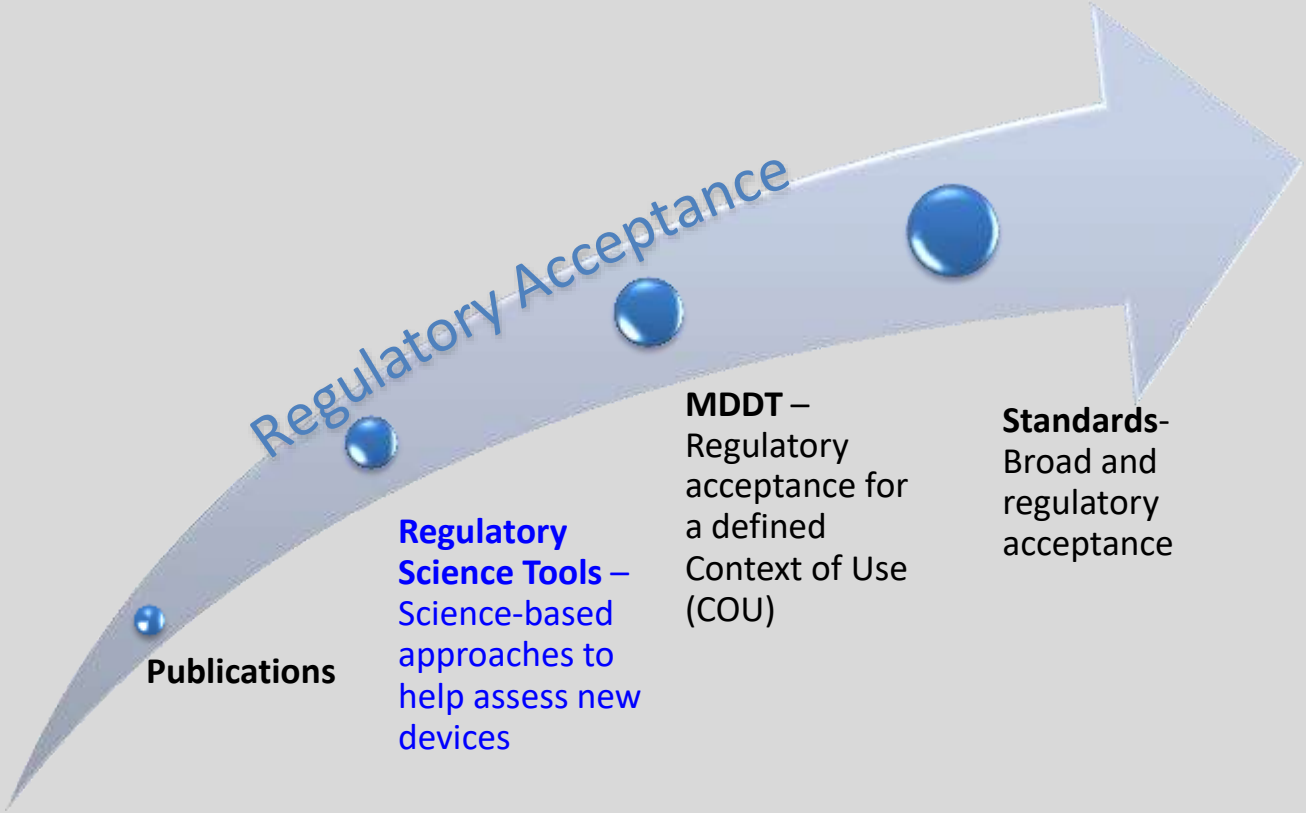
1. Computer aided diagnosis or detection AI systems (first CAD systems approved by FDA in 1990s)
2. Autonomous AI systems for patient referral
3. Systems for patient triage, rule-in, or rule-out

Putting Tools in Hands of Stakeholders

Regulatory Science Tools (RST)



Regulatory Science Tools (RST)



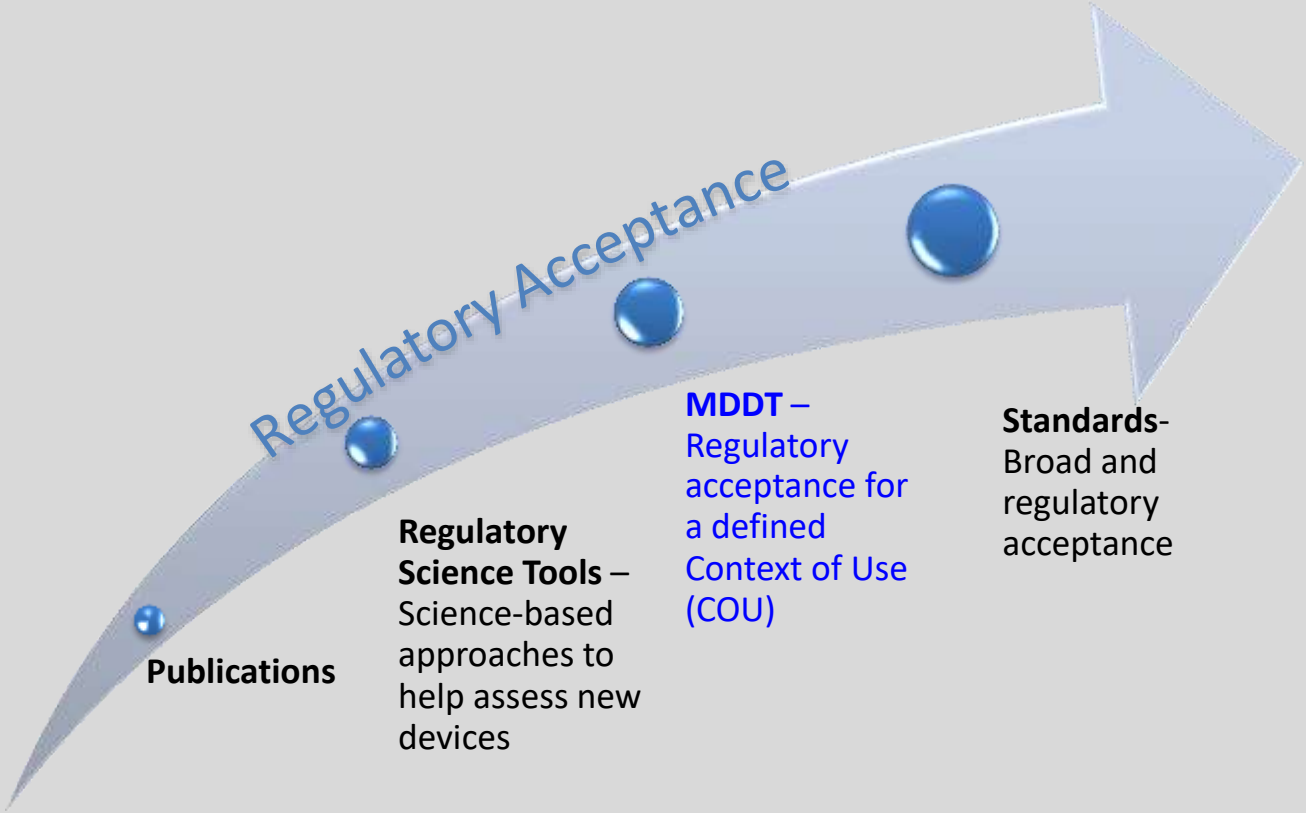
AI/ML Relevant RSTs

iMRMC: Multi-Reader Multi-Case Reader Studies	Statistical tools (java GUI and R package) to analyze, size, and simulate multi-reader multi-case (MRMC) reader studies	Model	Imaging reader studies, Artificial intelligence/machine learning	GitHub
iRoeMetz Application	A java application used to simulate reader scores for multi-reader multi-case (MRMC) reader studies	Model	Imaging reader studies, Artificial intelligence/machine learning	GitHub

www.fda.gov/medical-devices/science-and-research-medical-devices/catalog-regulatory-science-tools-help-assess-new-medical-devices

VICTRE: Breast Mass Generation Software	A modeling software that randomly generates main body of breast masses including random branching spicules grown out from the mass surface	Model	Medical imaging and diagnostics	GitHub
VICTRE: Digital Mammography Regions of Interest (ROIs)	VICTRE ROI patches for digital mammography of breast density categories with microcalcification cluster and spiculated mass inserted signals.	Dataset	Medical imaging and diagnostics	GitHub
VICTRE: Model Observers (MO)	Computer model observer functions to perform location-known lesion detection tasks	Model	Medical imaging and diagnostics	GitHub
VICTRE: Virtual Imaging Clinical Trials for Regulatory Evaluation	An entirely in-silico imaging clinical trial replicating a premarket study.	Model	Medical imaging and diagnostics	GitHub Article
VICTRE: Multi-modality Anthropomorphic Breast Phantom	A digital breast phantom with modifiable parameters including phantom voxel size (resolution) and breast density	Phantom, Virtual	Medical imaging and diagnostics	GitHub Document
VICTRE_MCGPU: Pivotal Study Simulations	A simulation tool that replicates a Siemens Mammomat Inspiration system for VICTRE	Model	Medical imaging and diagnostics	GitHub

Regulatory Science Tools (RST)



Medical Device Development Tools (MDDTs)

Qualification of Medical Device Development Tools

Guidance for Industry, Tool Developers, and Food and Drug Administration Staff

Document issued on: August 10, 2017

The draft of this guidance document was issued on November 14, 2013.

For questions regarding this document, contact MDDT@fda.hhs.gov.

www.fda.gov/regulatory-information/search-fda-guidance-documents/qualification-medical-device-development-tools

Summary

- Active research from OSEL has been
  - Identifying and addressing critical gaps in device evaluation of medical AI/ML
  - Putting methodology and tools into the hands of stakeholders

Acknowledgments

- I'd like to acknowledge Berkman Sahiner, Nicholas Petrick, Brandon Nelson, Elena Sizikova, Kenny Cha, and Elim Thompson for providing slides and information used this presentation.

Questions