

Is this AI output reliable?

Towards trustworthy AI for medical imaging: Guidelines for evaluation

Abhinav K. Jha, PhD

Department of Biomedical Engineering

Mallinckrodt Institute of Radiology

SNMMI AI Taskforce Evaluation Team lead

CSE 531A Attendance Form



Artificial intelligence (AI) is showing significant promise in multiple areas!



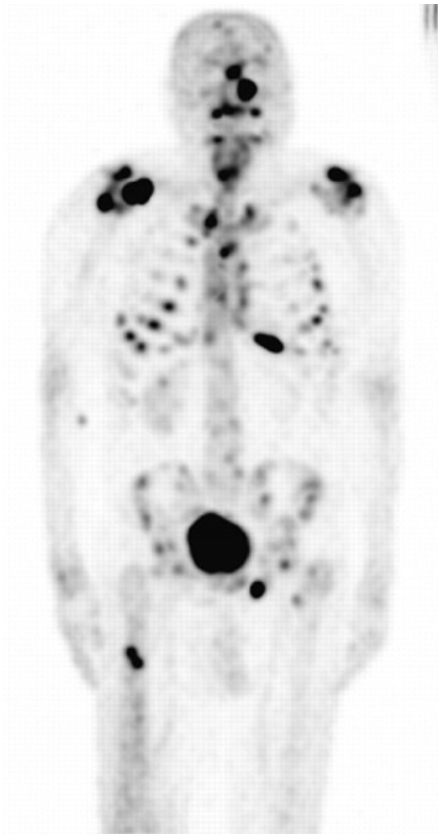


*Image source: https://www.javatpoint.com/application-of-ai

*Source: Pandya et al, Sensors, 2021

The broad application in this talk: Medical Imaging





The technique of creating images of the interior of a body for clinical analysis

Nuclear-medicine imaging: PET and SPECT



Single-photon emission computed tomography (SPECT)



Positron emission tomography (PET)

Imaging modalities that are widely used in multiple clinical applications

The role of computation in nuclear-medicine imaging



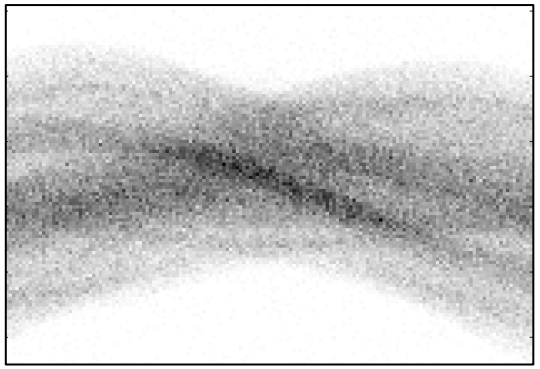


Image reconstruction

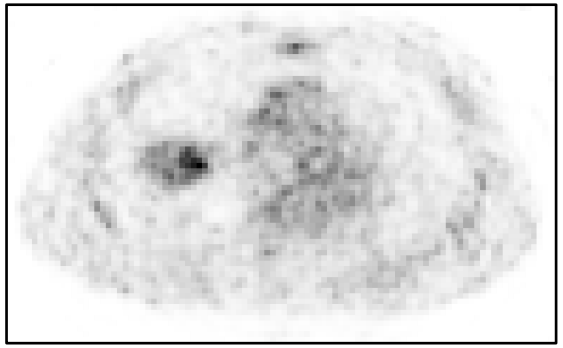


Image enhancement

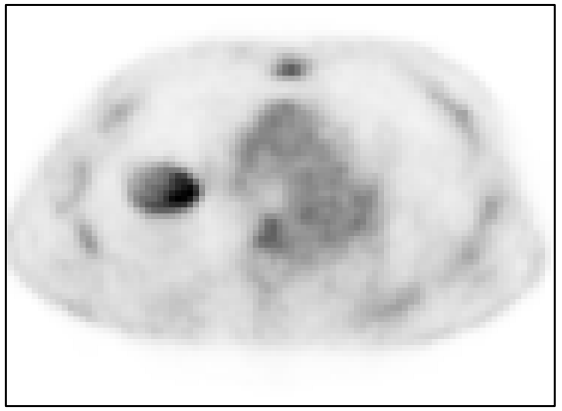
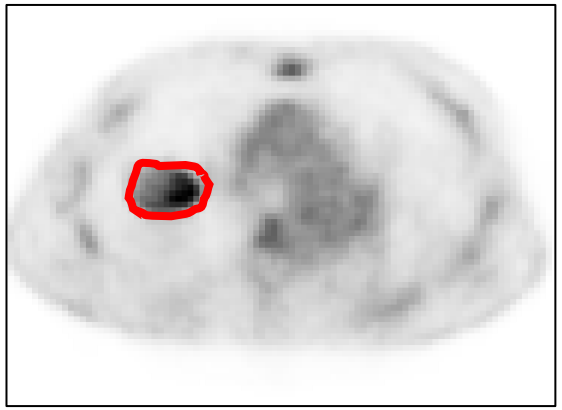
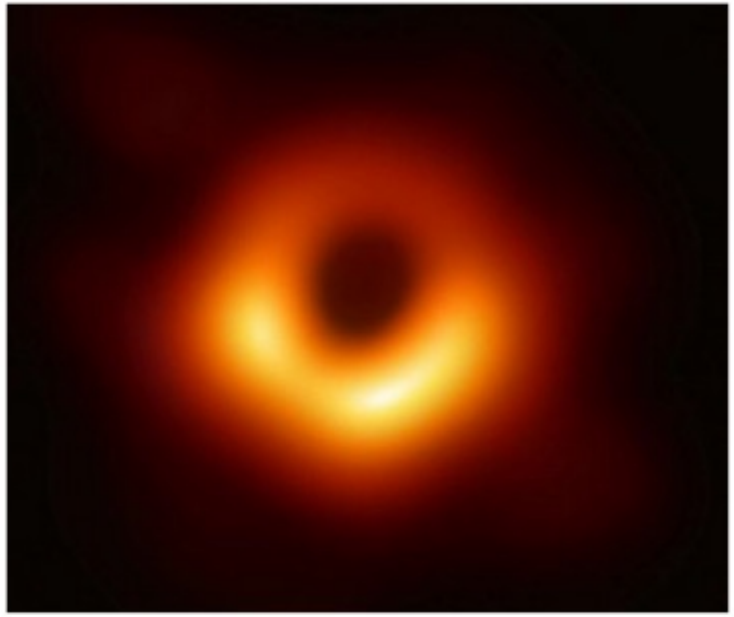


Image segmentation



Parameter quantification and clinical decision making

The role of computation in nuclear-medicine imaging



Black hole: Heart of the M87!



Image of a human heart with SPECT!

*Image source: Fortney

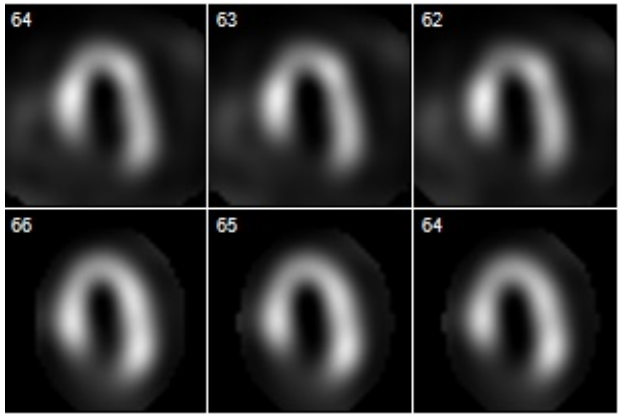
The specific clinical application for this talk: Cardiac SPECT



Patients with symptoms of coronary artery disease

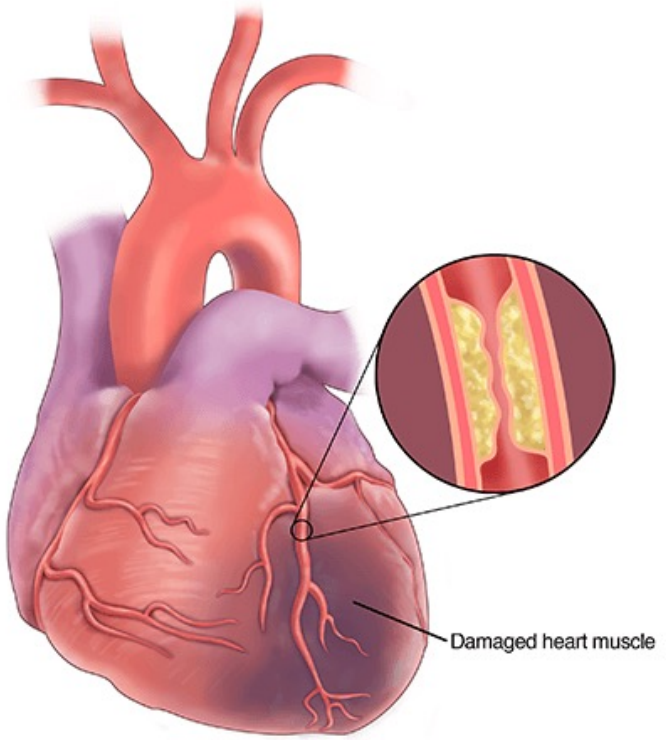


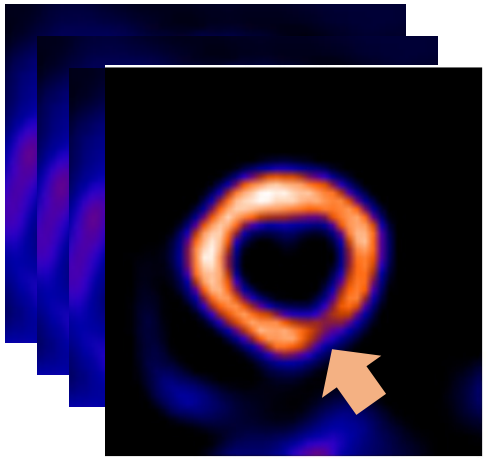
SPECT scanner



Cardiac SPECT images

Clinical application: Cardiac SPECT



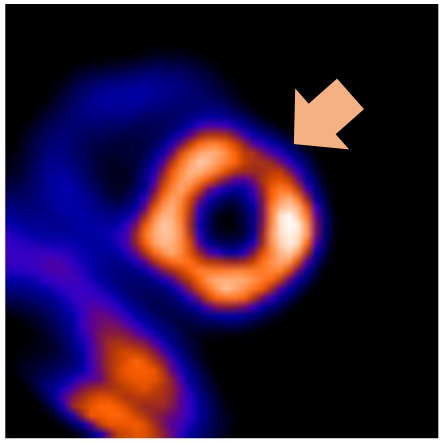


Area of reduced blood flow (defect)

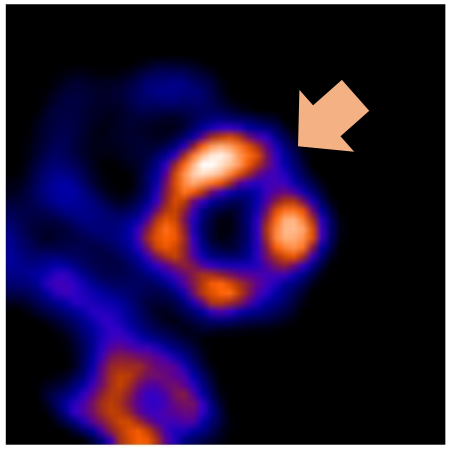
*Source: upmc.com

Motivating application

- Cardiac SPECT procedures require:
  - Patient to be administered radiation dose (~48 mCi)
  - Be scanned for a long time on the scanner
- Important need to reduce radiation dose/acquisition time for patients, but that impacts image quality.



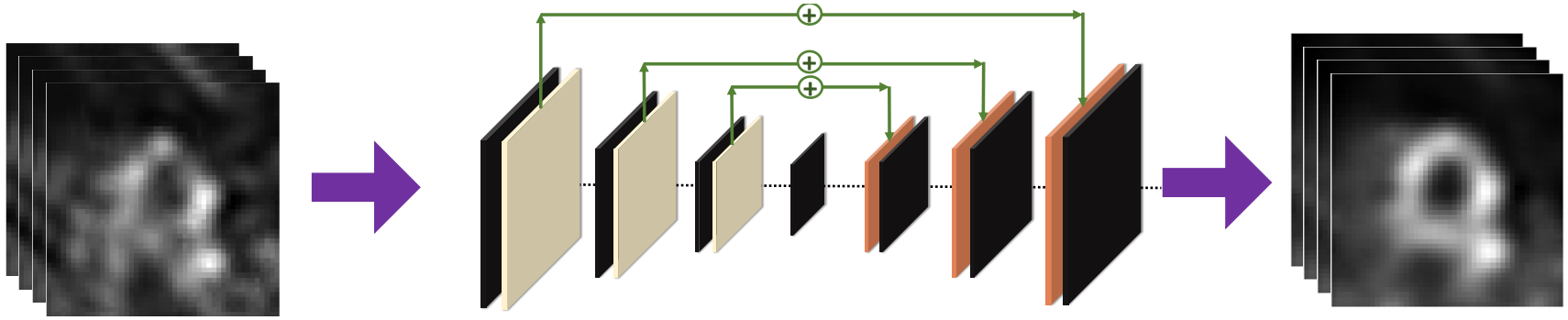
Normal-dose image



Low-dose image

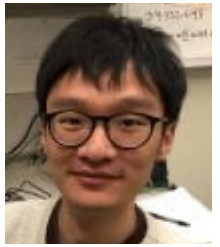
The role of AI

- Showing promise in predicting normal-dose images from low-dose images



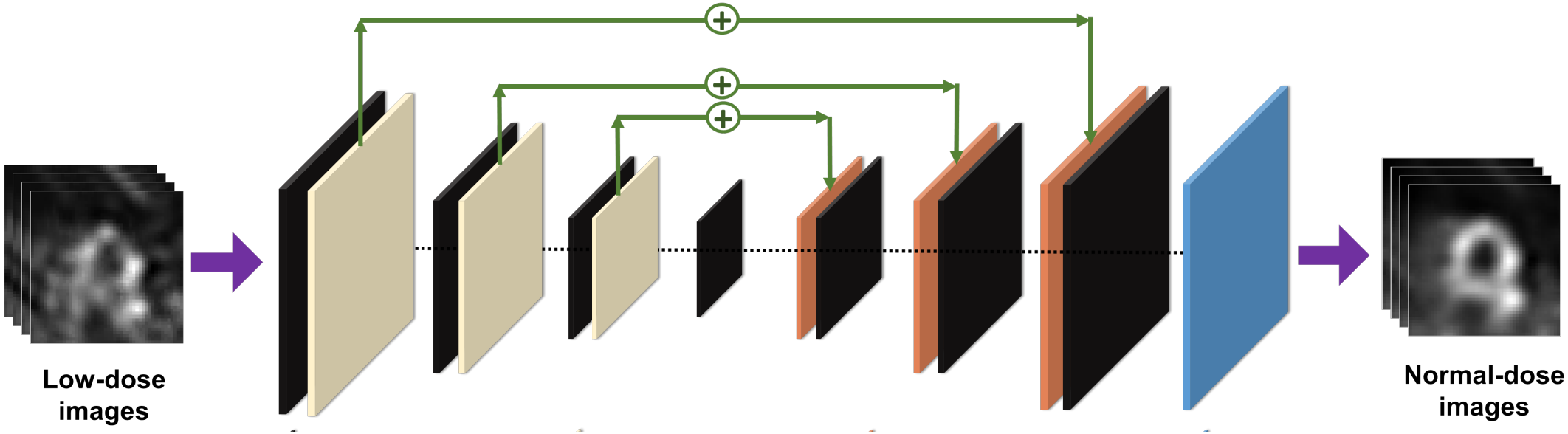
- For translation of these algorithms, rigorous evaluation is crucial
- AI algorithms learn from analysis of training data. Thus:
  - Their performance depends heavily on the training data
  - Output often not interpretable and can be unpredictable
- This leads to unique challenges while designing evaluation strategies

Task-agnostic evaluation can be misleading



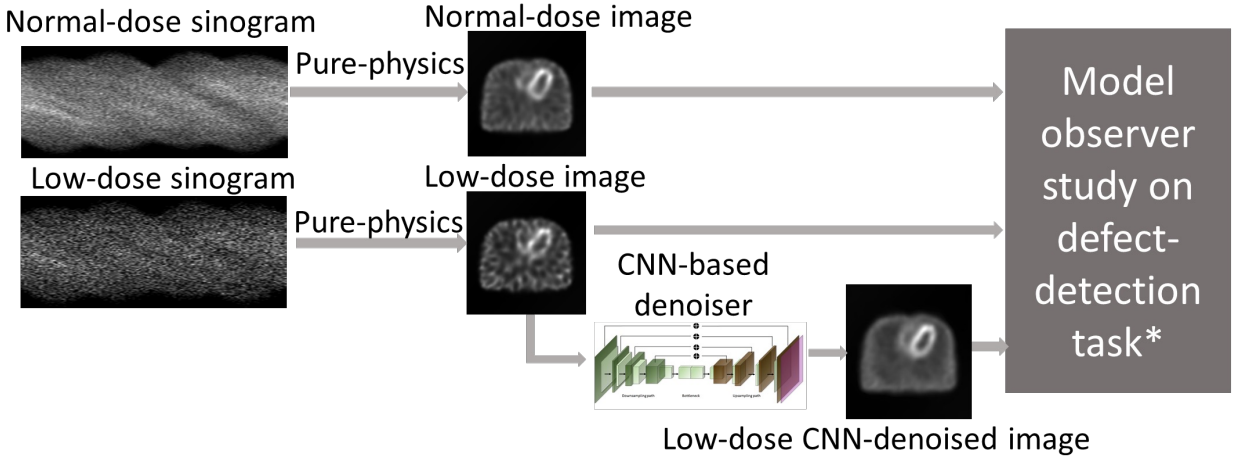
Zitong Yu

Context: Evaluating an AI-based “denoising” technique developed to reduce the injected radiation dose/acquisition time for myocardial perfusion SPECT

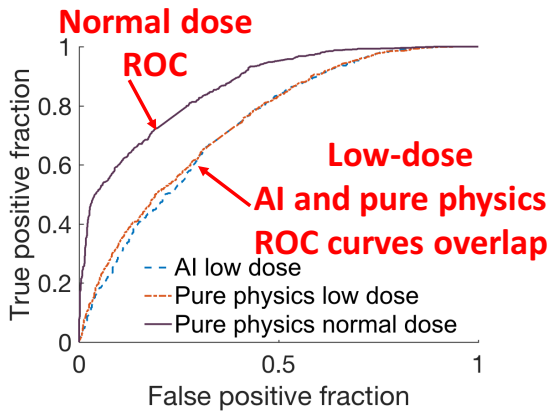


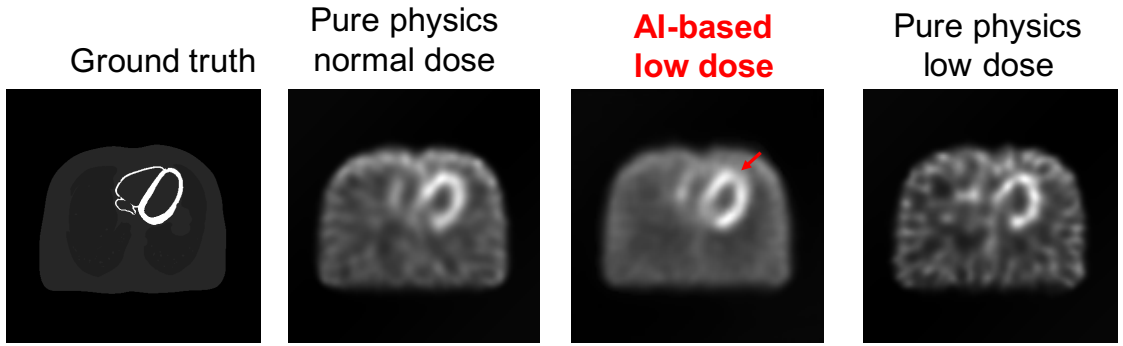
Conventionally, these techniques are evaluated using metrics that measure the fidelity between the AI-predicted normal-dose and the true normal-dose image

Task-agnostic evaluation can be misleading



	Conventional metrics			Task-specific metric
	Root mean square error	Structural similarity index	Peak signal-to-noise ratio	Area under the ROC curve (AUC)
AI-based	0.045 (0.045,0.046)	0.87 (0.87,0.87)	27.25 (27.13,27.35)	0.74 (0.71,0.76)
Pure physics-based	0.047 (0.047,0.048)	0.85 (0.85,0.85)	26.80 (26.69,26.90)	0.74 (0.72,0.77)
Inference	**AI better ($p < 0.01$)**	**AI better ($p < 0.01$)**	**AI better ($p < 0.01$)**	**AI gives similar performance as physics!**





Evaluation with conventional metrics may provide misleading interpretations

*Yu et al, Med. Phys. 2023

Need for task-specific evaluation

Work highlighted in media

eMedNews
Health News on the go.

HEADLINES Type 2 Diabetes at 30s can Reduce Life Expectancy

Home COVID-19 Research Health and Wellness Medical Innovations CME/CE Conferences About Us

Medical Innovations

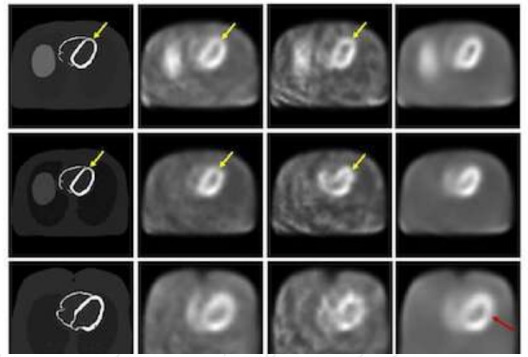
Medical Imaging with AI, Clinical Utility, Not 'Prettiness'

By Haritha Karyampudi on June 20, 2023

JUNE 19, 2023 Editors' notes

Clinical utility, not 'prettiness,' best metric for evaluating AI improvements to medical imaging, say engineers

by Shawn Ballard, Washington University in St. Louis



RECENT POSTS

TechTarget HEALTH IT ANALYTICS xtelligent HEALTHCARE MEDIA

Home News Features Interviews P

Population Health Precision Medicine Quality & Governance Tools & Strategies Focus on Population Health

TOOLS & STRATEGIES NEWS

AI-Based "Denoising" in Medical Imaging May Require Further Evaluation

Researchers are calling for objective evaluations of deep learning-based tools to clean up medical images to ensure that these actually improve performance on clinical tasks.

healthcare-in-europe.com

News • Performance-based evaluation

Useful, not pretty: How AI can improve medical imaging

Medical imaging plays an essential role in diagnosis and treatment for an array of conditions. From X-rays to see a broken bone or a tooth cavity to SPECT scans for spotting heart defects, doctors use medical imaging to look inside the body, find disease and treat it appropriately. But what happens when those images aren't clear?

Challenge: Generalizability



Trained at this center and worked great!





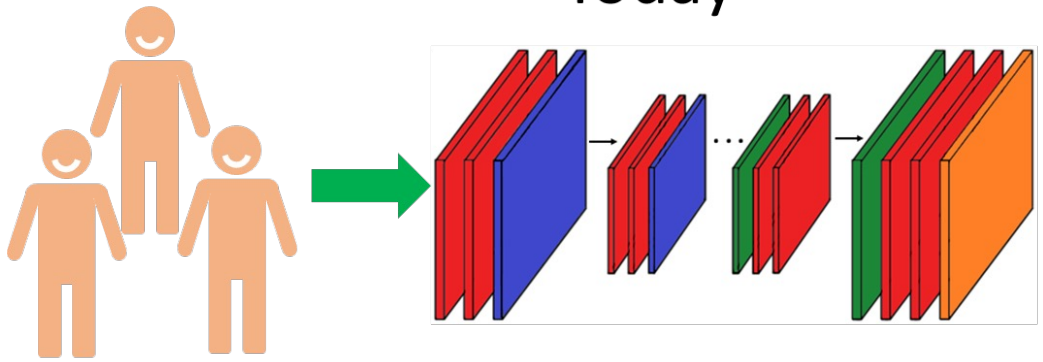
Oops!

Evaluation should characterize the generalizability of AI methods

Zech et al, PLOS Med, 2018
Gianfranceso et al, JAMA Intern Med. 2018
Noor et al, BMJ 2020

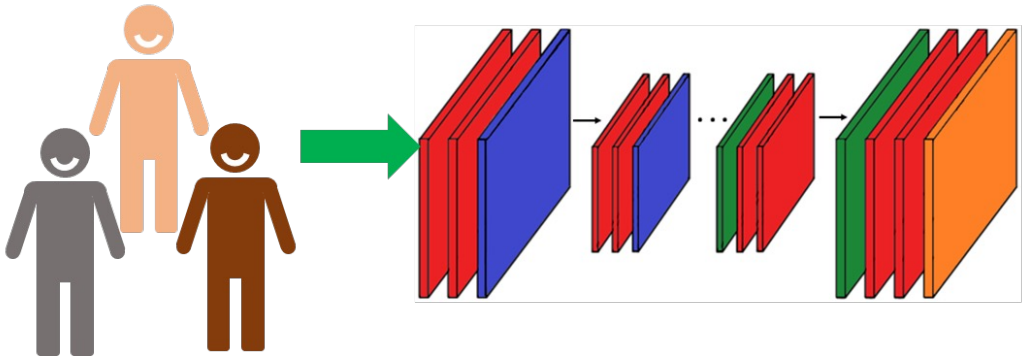
Challenge: Data drift

Today



Works great!

5 years later



May not work as well!

Evaluation should assess if the method is performing reliably in a post-deployment setting

SNMMI AI Taskforce Evaluation Team



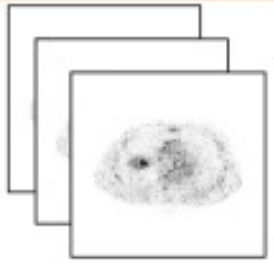
Team consisting of nuclear medicine physicists, computational imaging scientists, physicians, statisticians, representatives from the industry and from regulatory agencies

A key recommendation from the taskforce: The claim

An evaluation study for an AI algorithm should produce an accompanying claim consisting of the following components

1. Definition of clinical task

- Classify
- Quantify
- Jointly classify & quantify



2. Patient population

- Should be representative of target population
- Demographics including sex, age, ethnicity should be stated



3. Imaging process

- Imaging system(s)
- image-acquisition protocol(s)
- Single/multi center



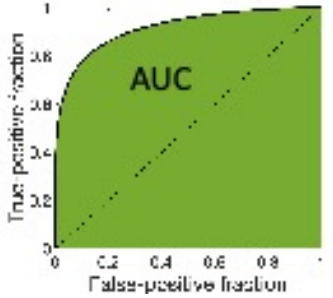
4. Strategy to extract task-specific information

- Detection: Human/model observers
- Quantification: Bayesian and frequentist
- Single/multi reader studies

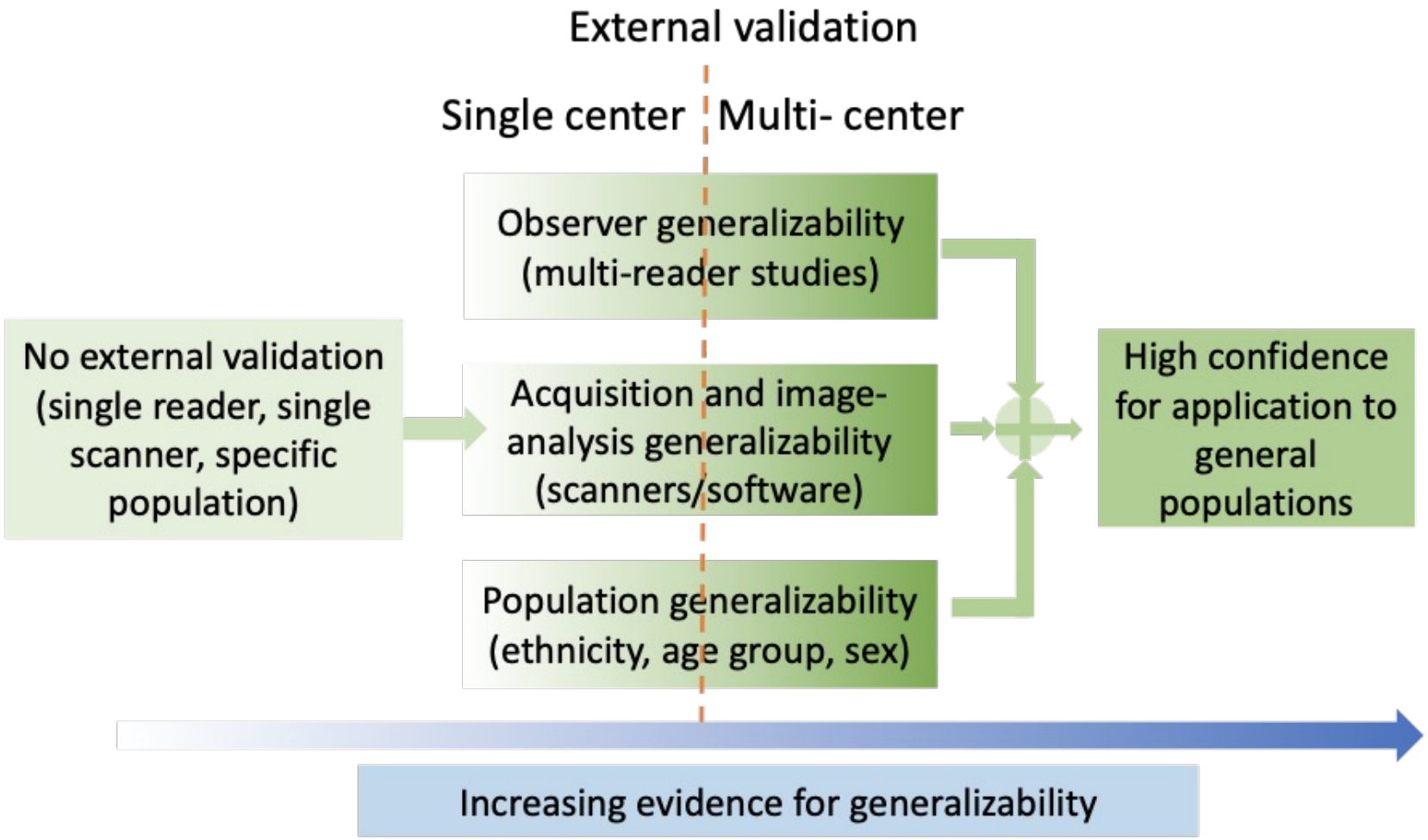


5. Figure of merit to quantify performance

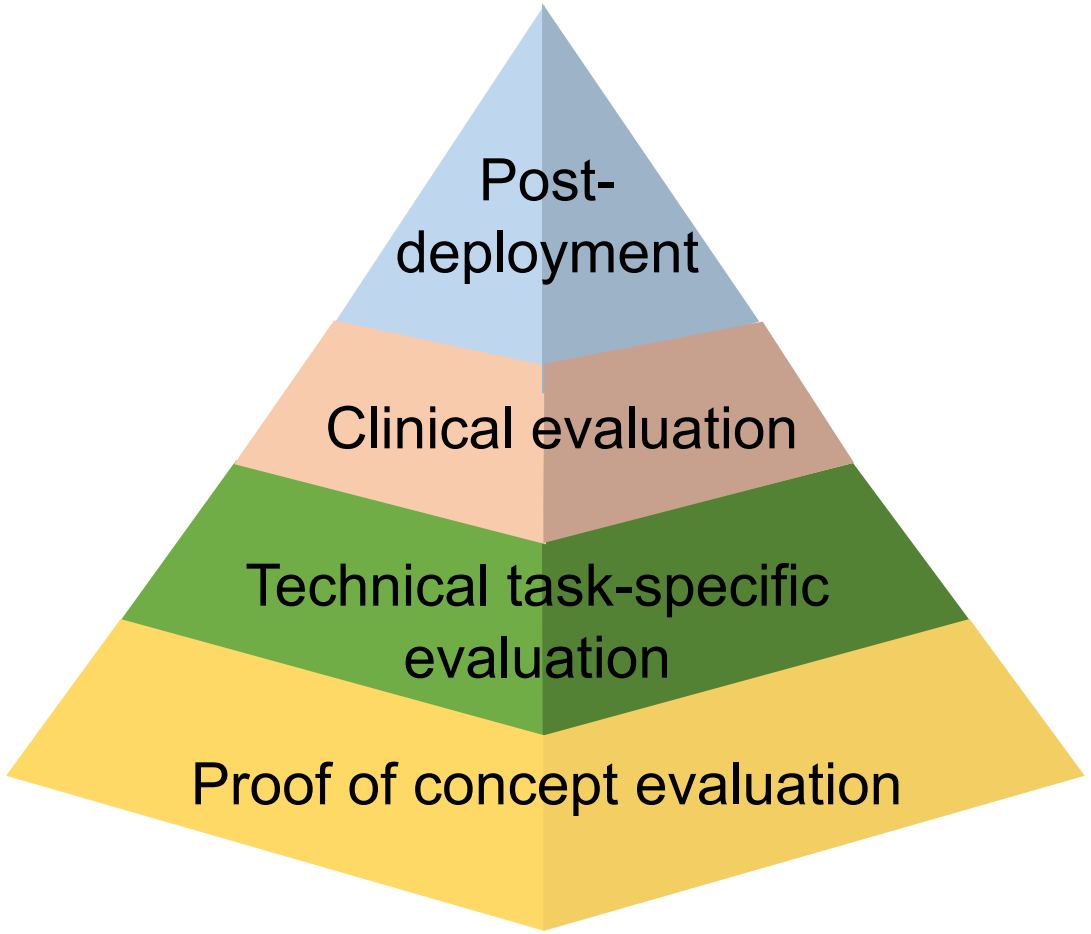
- Detection: ROC analysis
- Quantification: Ensemble bias and variance, EMSE, bias/variance profiles
- Jointly detect and quantify: LROC, EROC, FROC, LROC



The claim will inherently quantify the generalizability of the AI algorithm



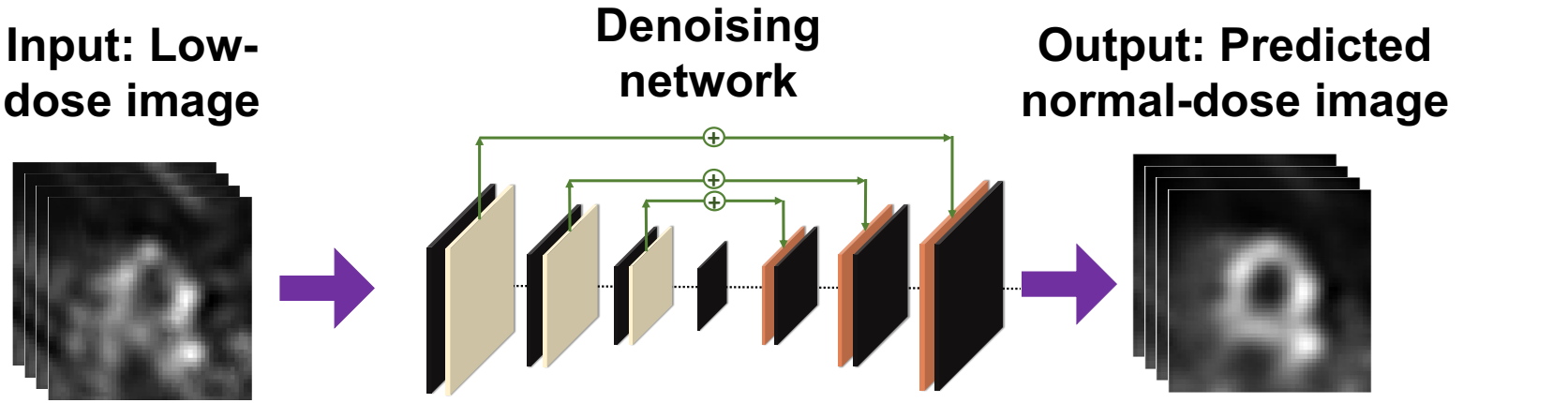
The task force proposed an evaluation framework



Task-based evaluation was a key theme underlying this framework

Class 1: Proof-of-concept evaluation

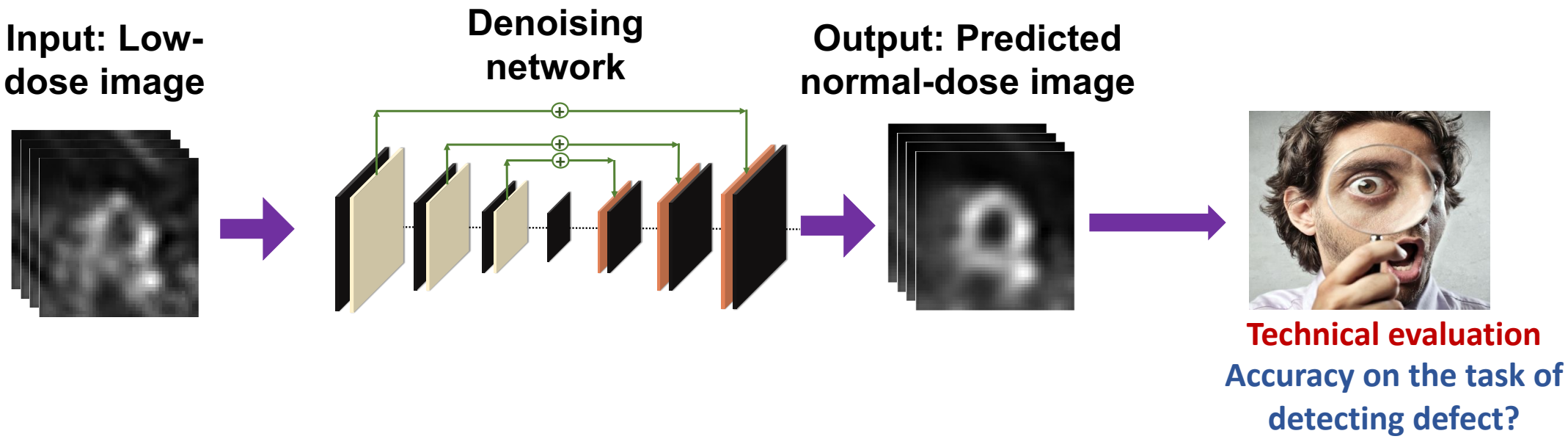
Objective: Demonstrate technical innovations of new methods



Proof of concept evaluation
Does predicted image match normal-dose image?

Class 2: Technical task-specific evaluation

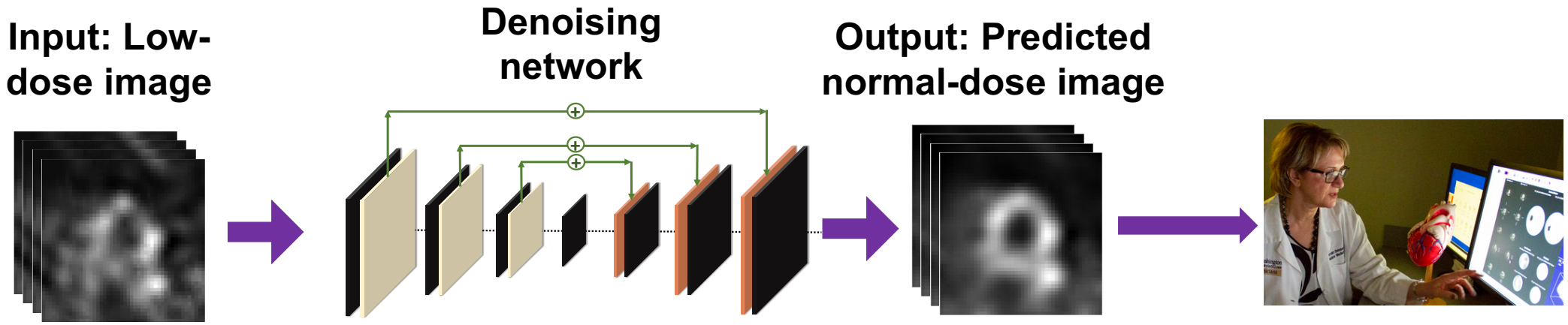
Objective: Quantify technical factors such as accuracy of the method on the specific clinical task



Example claim: An AI-based denoising method for MPI SPECT was superior to images reconstructed at low dose on the task of detecting perfusion defects with 80% power and a significance level of 5%. The AUC difference was within a pre-defined margin of 0.1.

Class 3: Clinical evaluation

Objective: Quantify efficacy of the method for making clinical decisions



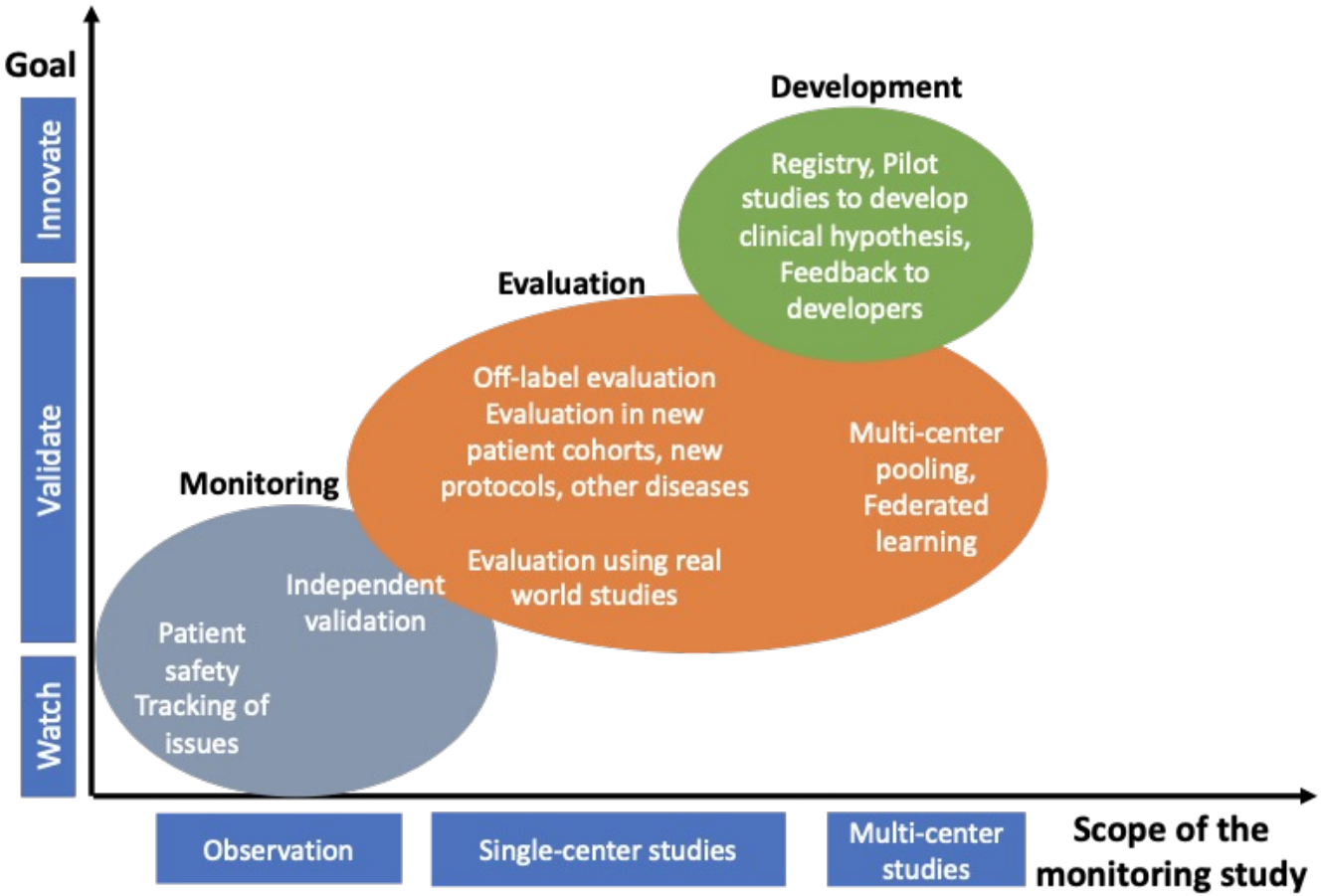
Clinical evaluation
Accuracy on the task of diagnosing CAD in clinical settings

Example claim: The average AUC of three experienced physicians on the task of diagnosing CAD by reading low-dose SPECT MPI images increased from X to Y (increase of ΔAUC (95% CIs)) when these images were reconstructed using an AI-based denoising method as compared to no denoising, as evaluated in a blinded retrospective study with clinical data collected from two institutions.

Class 4: Post-deployment evaluation

Objectives

- Monitoring, detecting technical issues, potential bugs, reportable events, opportunities for improvement
- Evaluating off-label use: Again task-based
- Provide feedback for development

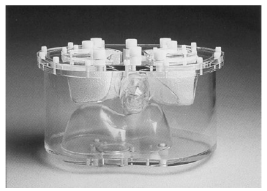


Elements of study design

1. Determine study type

- Simulations
- Phantoms
- Clinical studies
  - Retrospective/prospective
  - Observational/interventional





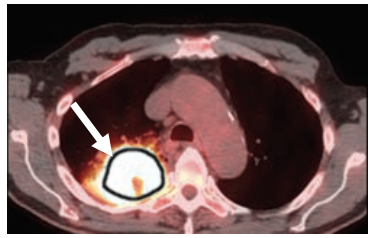
2. Data collection

- Representative
- No overlap between test and training data
- External validation for clinical/technical evaluation



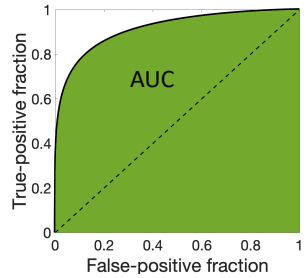
3. Define reference standard

- Quality may vary based on the claim
- Single/multi-reader studies
- Clinical outcomes



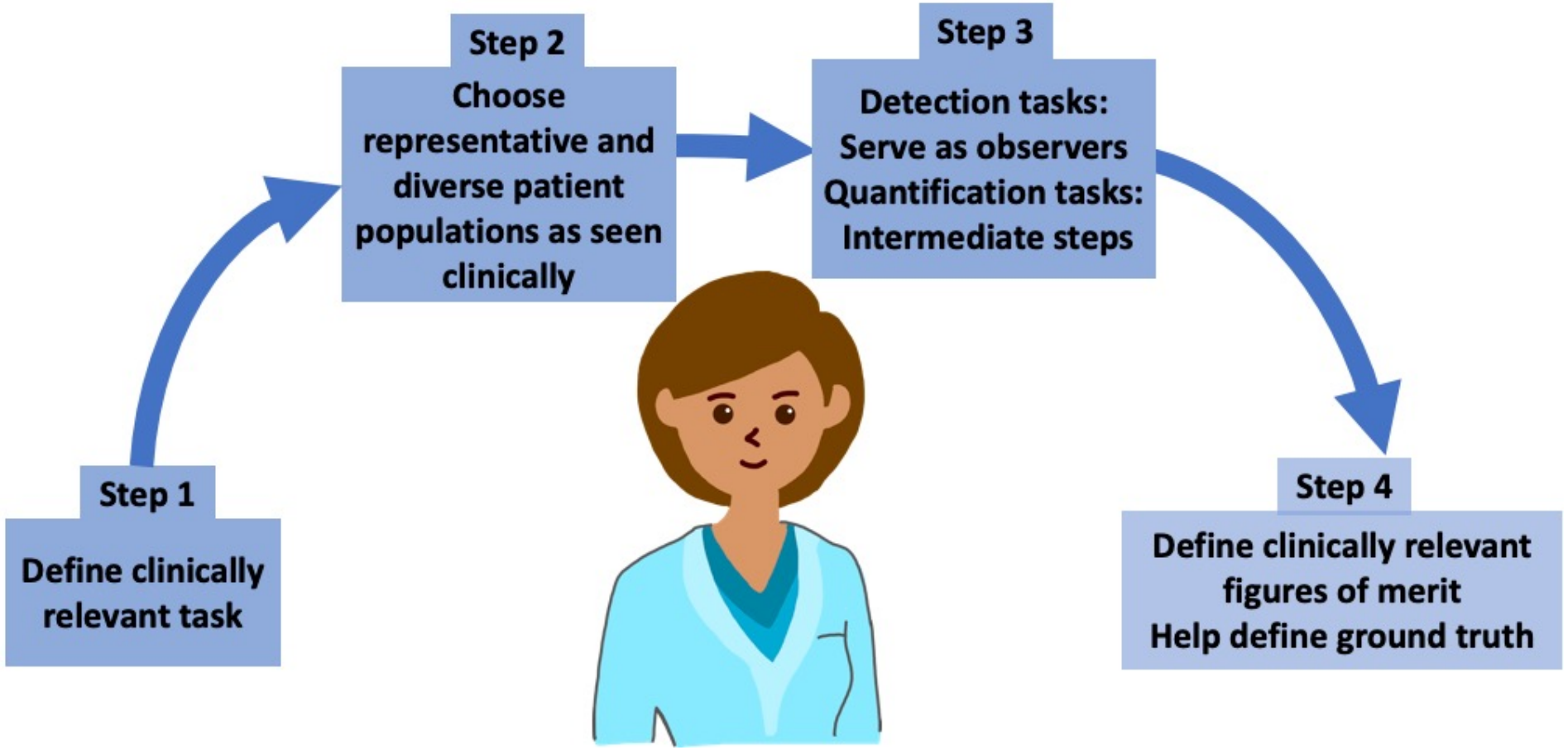
4. Choosing figure of merit

- Proof-of-concept eval.: Show clinical promise
- Technical eval.: Accuracy, Precision
- Clinical eval. Show clinical utility



The taskforce has provided the RELAINCE (**R**ecommendations for **E**va**l**uation of **AI** in **N**u**c**lear Medicin**e**) guidelines for each element of study design

End-users of AI tools have an important role in performing evaluations



*Jha et al, PET Clinics, 2021

The RELAINCE guidelines

Class of evaluation	Recommendation
Proof of concept	Ensure no overlap between development and testing cohort
	Check that ground-truth quality is reasonable.
	Provide comparison with conventional and state-of-the-art methods.
	Choose figures of merit that motivate further clinical evaluation.
Technical evaluation	Choose clinically relevant task: Detection/quantification/combination of both.
	Determine the right study type: Simulation/phantom/clinical.
	Ensure that simulation studies are realistic and account for population variability.
	Testing cohort should be external.
	Ensure that ground truth is high quality and represents task performance.
	Define an optimal strategy to extract task-specific information
	Choose figures of merit that quantify task performance.
Clinical evaluation	Determine study type: Retrospective, prospective observational, prospective interventional, or real-world studies
	Testing cohort must be external.
	Collected data should represent the target population as stated in the claim.
	Ground truth should be very high quality and be representative of clinical utility. Gold standard ground truth is clinical outcomes
	Figure of merit should reflect performance on clinical task
Post-deployment monitoring	Imperative to monitor devices and follow reporting guidelines
	Consider designing routine phantom studies as sanity checks to assess day-to-day performance
	Monitor data drift
	For off-label evaluation, follow recommendations as in clinical/technical evaluation depending on objective.

THE STATE OF THE ART

Nuclear Medicine and Artificial Intelligence: Best Practices for Evaluation (the RELAINCE Guidelines)

Abhinav K. Jha[1], Tyler J. Bradshaw[2], Irène Buvat[3], Mathieu Hatt[4], Prabhat KC[5], Chi Liu[6], Nancy F. Obuchowski[7], Babak Saboury[8], Piotr J. Slomka[9], John J. Sunderland[10], Richard L. Wahl[11], Zitong Yu[12], Sven Zuehlsdorff[13], Arman Rahmim[14], and Ronald Boellaard[15]

Work highlighted in media

AuntMinnie.com

Home | Webinars | Cases | CME | Communities | Conferences | Europe | Forums | Jobs

Best practices could help AI deployment in nuclear medicine

By Will Morton, AuntMinnie.com staff writer

May 31, 2022 -- Nuclear medicine imaging lacks a rigorous strategy f
artificial intelligence (AI) algorithms developed for clinical use, say an
group of experts. To that end, they recently offered a set of best pract
say could help transfer research into the field.

Evaluation of AI for medical imaging: A key requirement for clinical translation

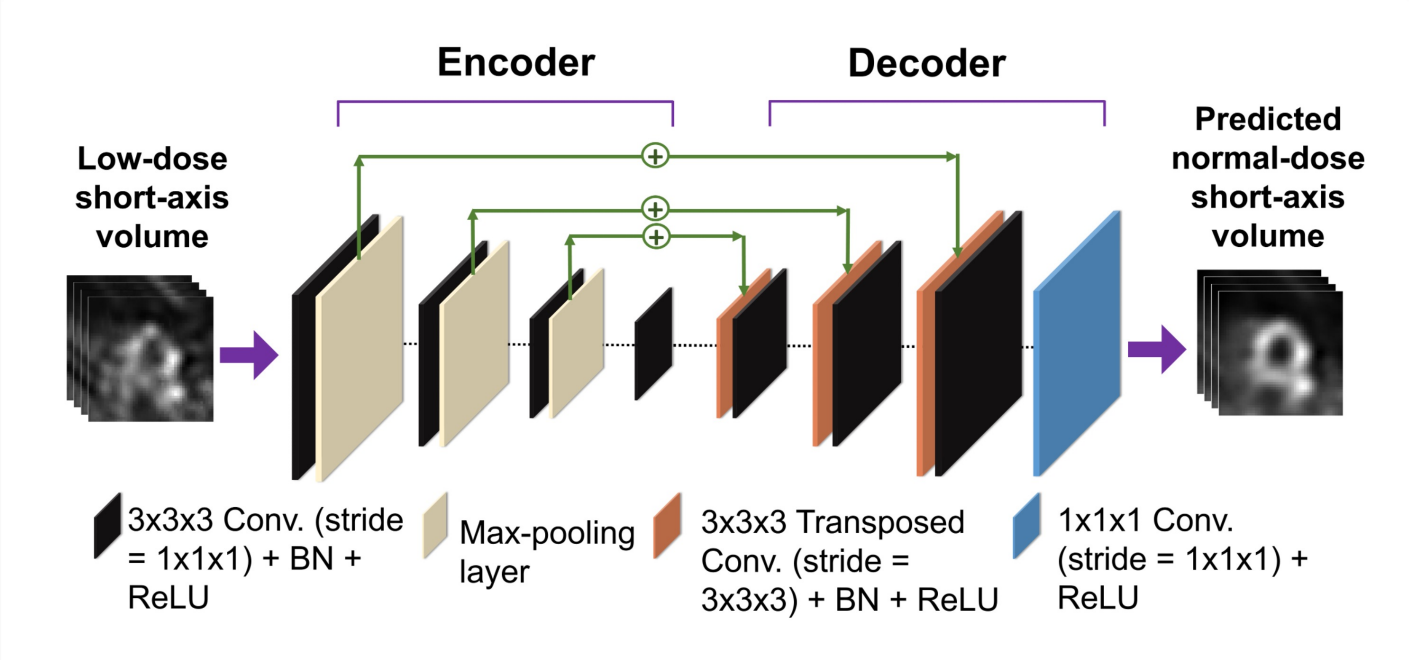
Task force Evaluation Team led by Abhinav Jha develops guidelines

Beth Miller 09.01.2022



To develop guidelines to evaluate artificial intelligence (AI) in nuclear-medicine imaging, an interdisciplinary team was set up within an AI taskforce established by Richard L. Wahl, MD, director of Mallinckrodt Institute of Radiology (MIR) at Washington University School of Medicine in St. Louis. The team, led by Abhinav Jha, assistant professor of biomedical engineering in the McKelvey School of Engineering and assistant professor of radiology at MIR, has published Recommendations for Evaluation of AI for Nuclear Medicine (RELAINCE) guidelines.

An example of using these guidelines to evaluate a novel AI method: DEMIST*





Ashequr Rahman

$$\text{Loss function} = \text{Fidelity loss} + \lambda * \text{Observer loss}$$

Minimize mean square error between normal-dose and predicted normal=dose image

Preserve features that assist in detection task: Minimize mean square error between channel features in normal and predicted normal-dose image

*Rahman et al, IEEE TRPMS April 2024

Evaluation: Patient data collection and curation

$N = 4118$ cases with MPI studies

$N = 1295$ cases with SPECT projection data, CT scans

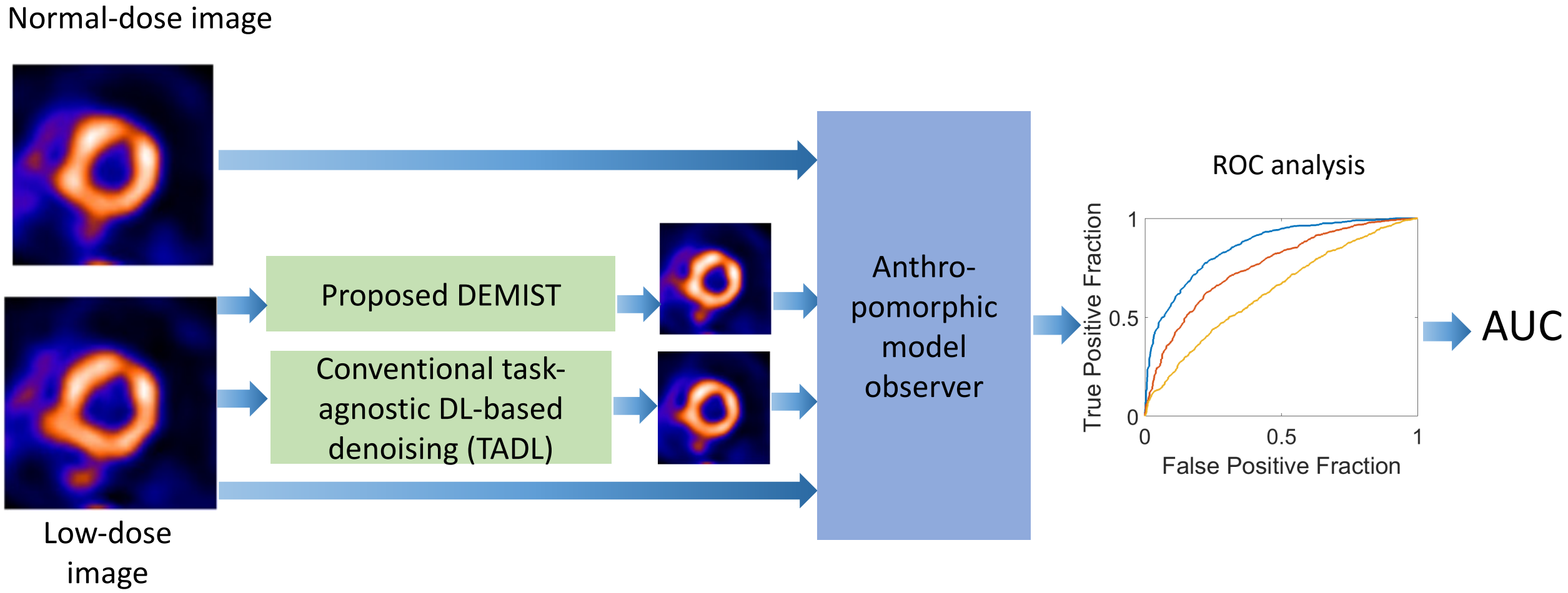
Excluded 957 cases
(diseased case, failed access,
post-processing errors, missing reports)

$N = 338$ normal cases (224 males, 111 females, 3 not reported)
Data acquired from two scanners (GE Discovery 670 Pro NaI and Pro CZT)

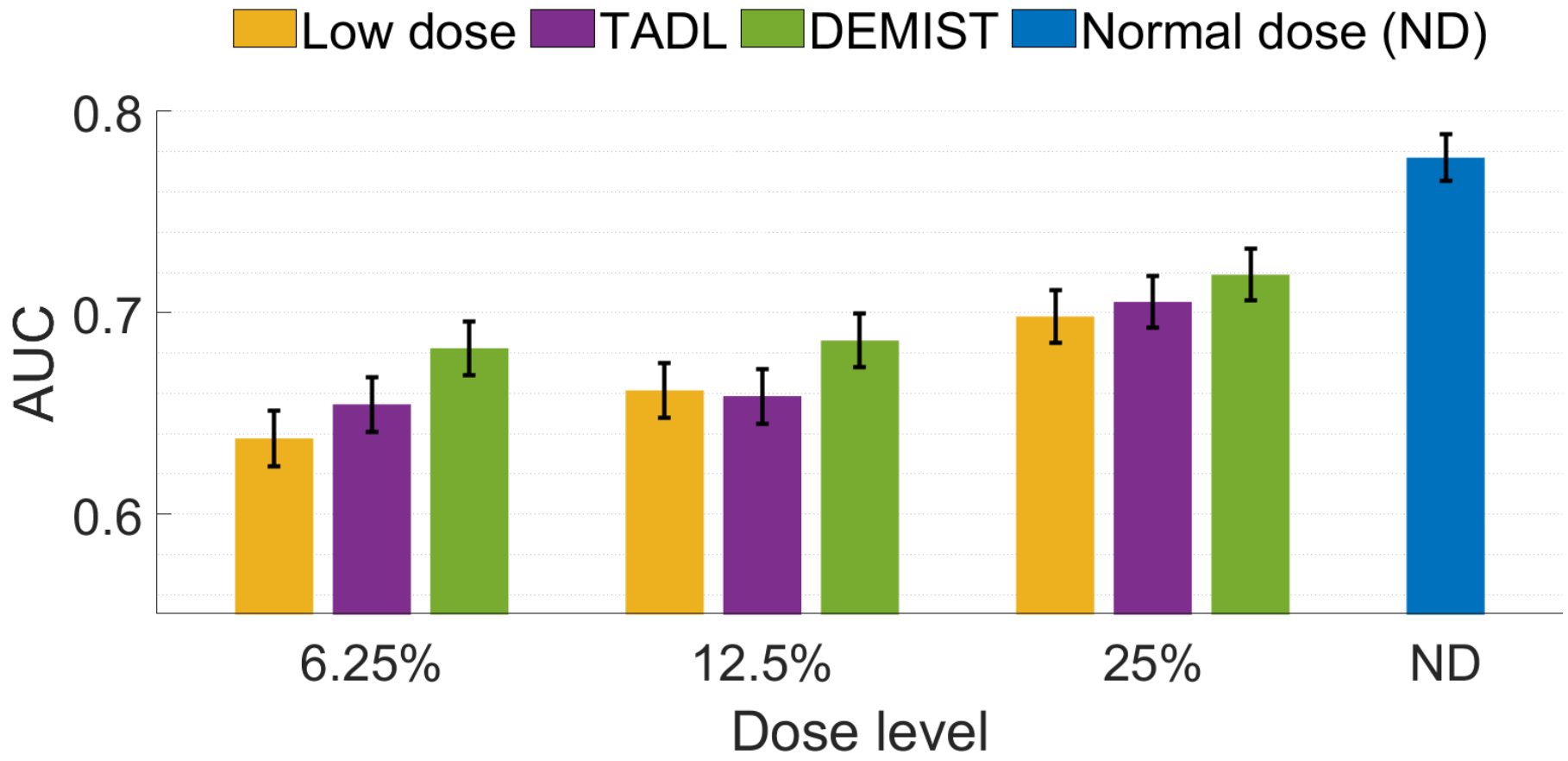
Training $N = 184$ normal cases		**Validation** $N = 40$ normal cases		**Test** $N = 114$ normal cases	
$N = 184\times4$ samples with inserted defects	$N = 184\times12$ samples without defects	$N = 20\times12$ samples with inserted defects	$N = 20\times12$ samples without defects	$N = 53\times18$ samples with inserted defects	$N = 61\times18$ samples without defects



Evaluation: Methodology to perform observer study

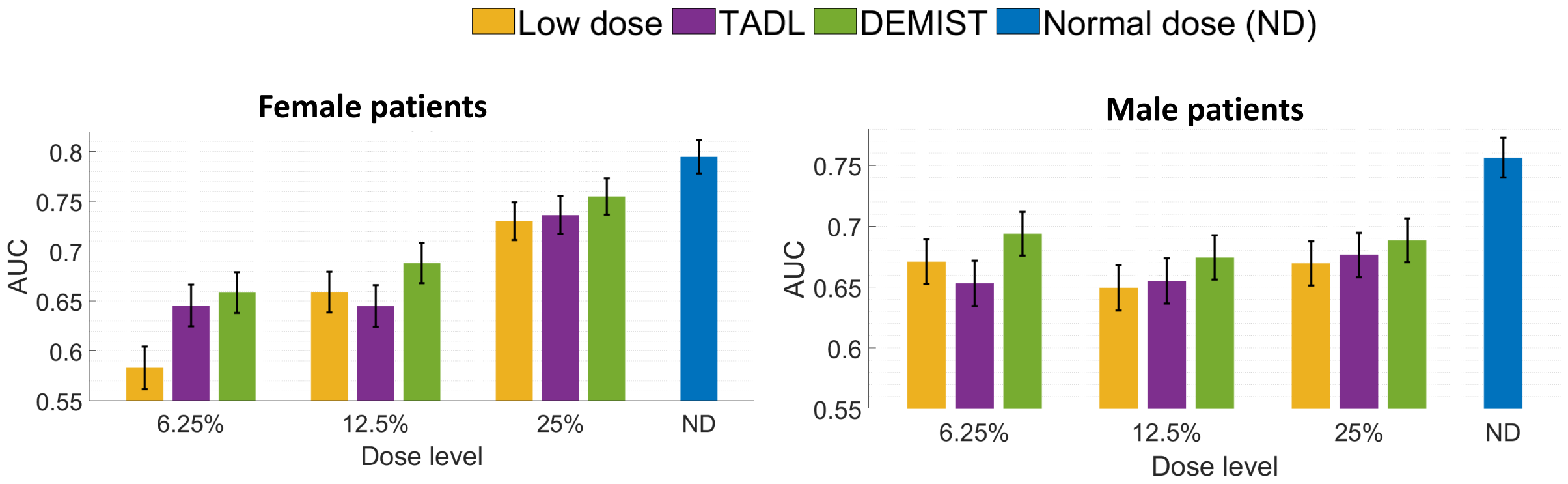


Results: Observer-study-based evaluation



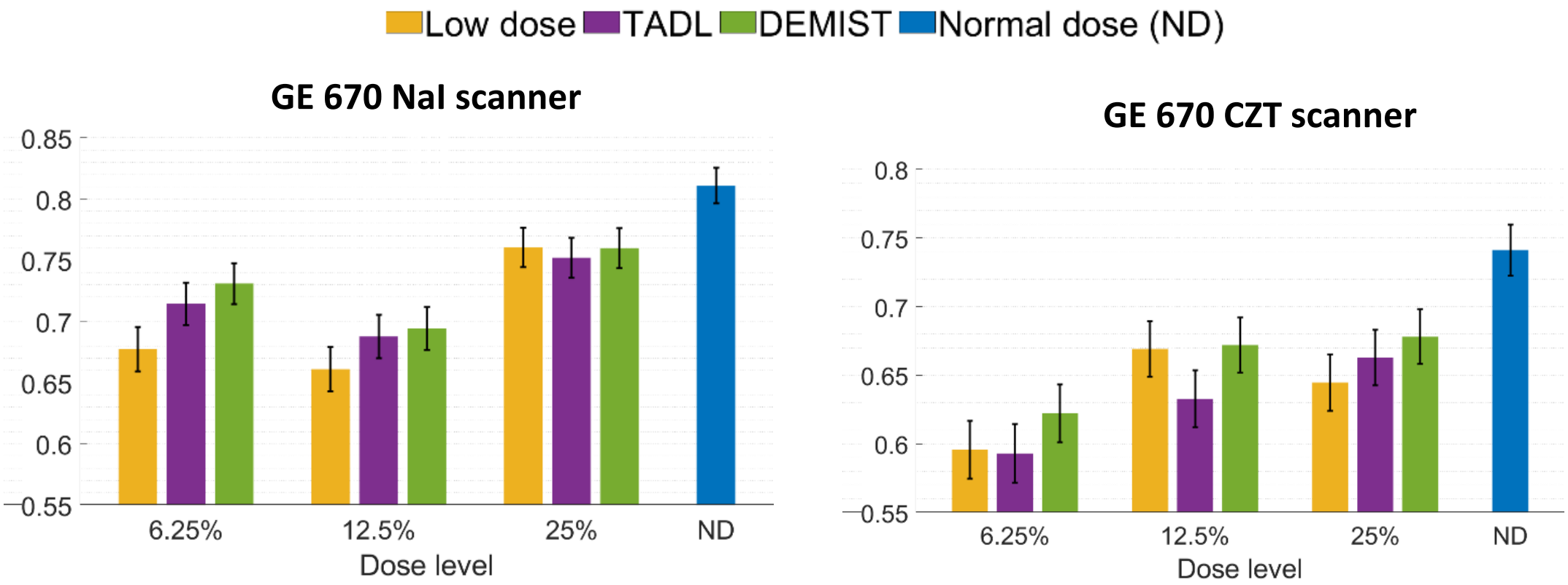
DEMIST significantly improved performance on defect detection task compared to low-dose protocol and TADL method.

Stratified analysis: Patient sex



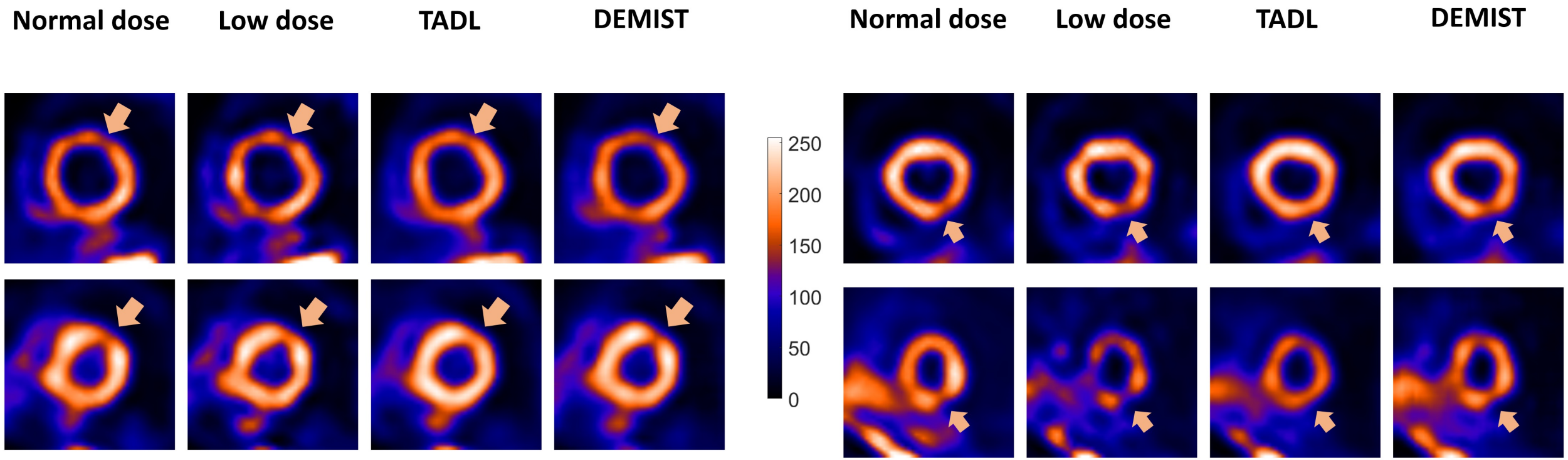
DEMIST significantly improved performance on defect-detection task compared to low-dose protocol for both sexes

Generalizability to different scanners



DEMIST significantly improved performance on defect-detection task compared to low-dose protocol for data from both scanners

Qualitative evaluation



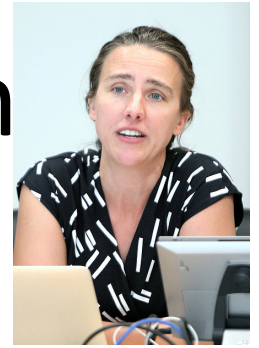
TADL method tends to wash out the defect. The proposed DEMIST method was able to preserve the defect features.

Claim for the proposed method

An AI-based denoising method for myocardial perfusion SPECT yielded statistically significant improved performance compared to the low-dose images on the task of detecting perfusion defects as evaluated using a retrospective clinical study with single-center data and with an anthropomorphic model observer.

*Rahman et al, IEEE TRPMS 2024

Important to be mindful of ethics as we push boundaries with AI



Anya Plutynski, PhD

What do patients want?

- Uncertainty of AI algorithms should be accounted for when making clinical decisions
- Patient risk-value profiles should be considered when using AI for making clinical decisions

nature medicine

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature medicine › comment › article

Comment | Published: 31 January 2022

Clinical decisions using AI must consider patient values

Jonathan Birch ✉, Kathleen A. Creel, Abhinav K. Jha & Anya Plutynski

Nature Medicine **28**, 229–232 (2022) | Cite this article

1038 Accesses | **30** Altmetric | Metrics

Built-in decision thresholds for AI diagnostics are ethically problematic, as patients may differ in their attitudes about the risk of false-positive and false-negative results, which will require that clinicians assess patient values.

Patients want AI in medicine, but to help physicians!

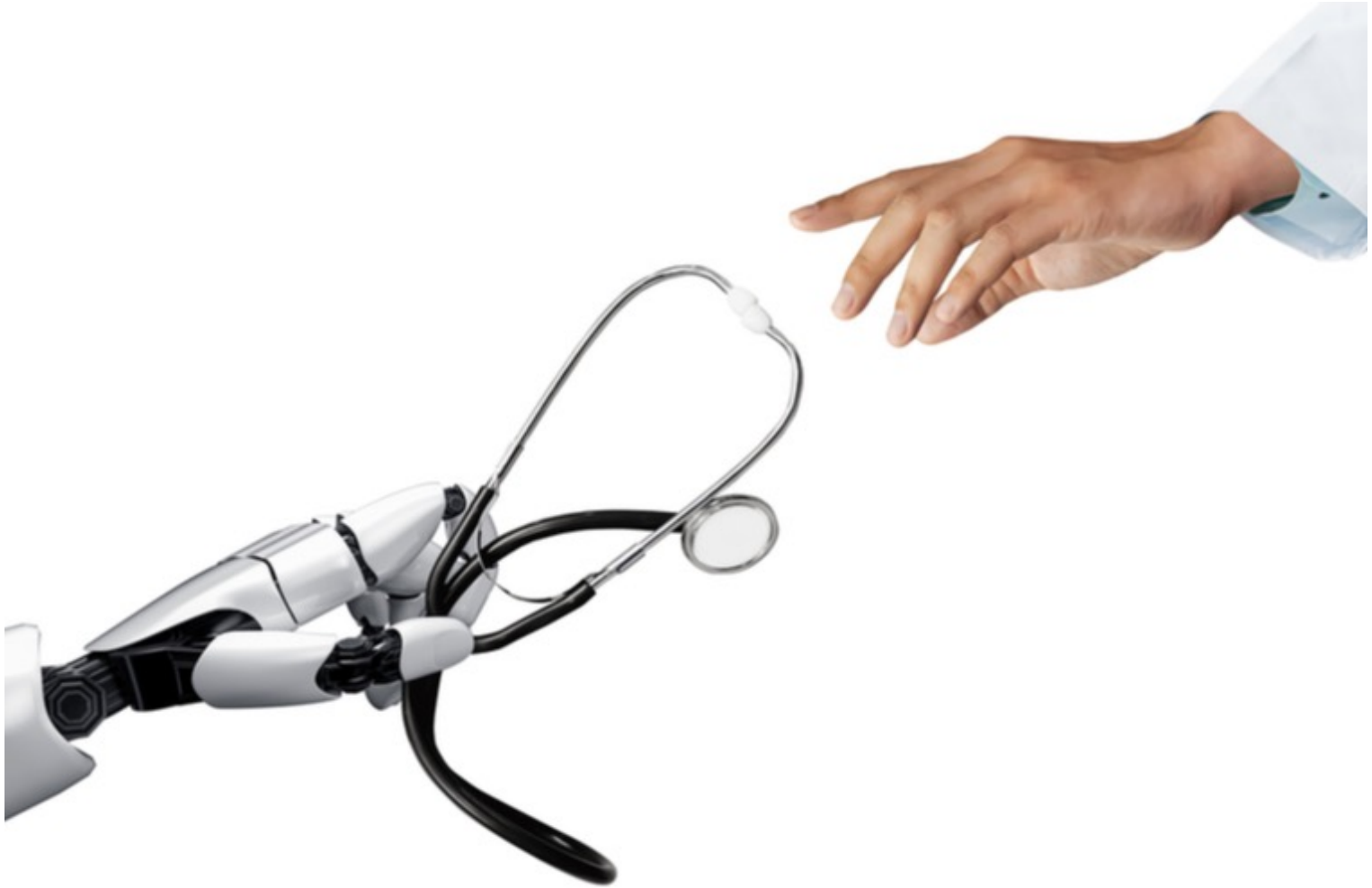
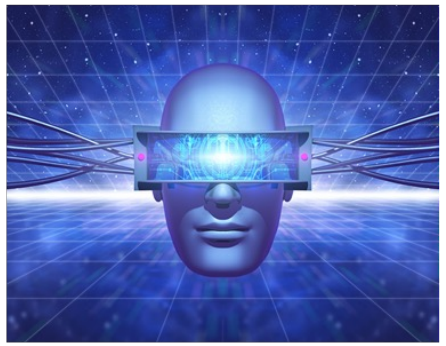


*Photo credit: Shutterstock

Conclusion

AI algorithms



+

Clinical task-based evaluation



=

Transform medical imaging



Acknowledgements



SNMMI AI Task Force

Financial support: NIH/NIBIB R01 EB031051, R01 EB031962, R56 EB028287, R21 EB024647 (Trailblazer award), R01 EB031051-S2, CHPC, NVIDIA GPU grant

Thank you!

Welcome questions and collaborations!



jhalab.wustl.edu



a.jha@wustl.edu



jhalab_washu