AI-based Quantitative Imaging Biomarkers

Robert Jeraj

Professor of Medical Physics, Human Oncology,

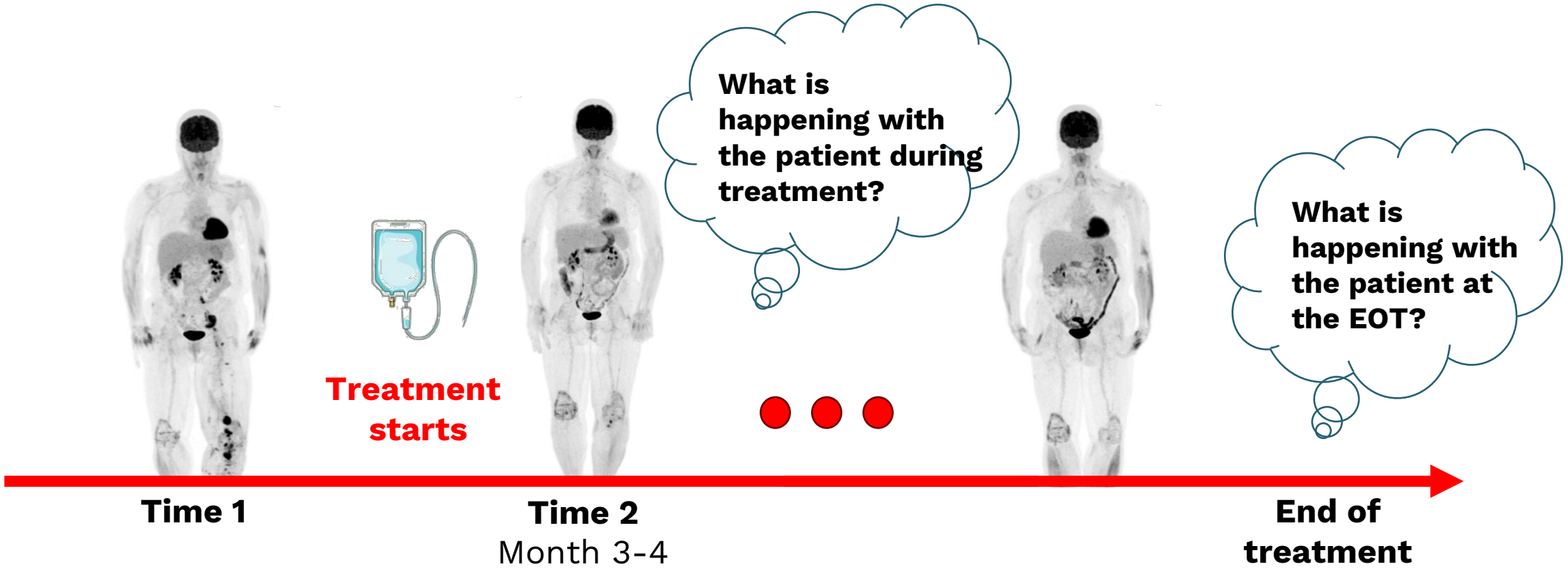
Radiology and Biomedical Engineering

University of Wisconsin, Madison, WI, USA

University of Ljubljana, Slovenia

rjeraj@wisc.edu

Patient journey



Biomarkers:
A key to the personalized cancer care

"The emerging use of **cancer biomarkers** may herald an era in which physicians no longer make treatment choices that are based on population-based statistics but rather on the **specific characteristics of individual patients and their tumor**."

Dalton and Friend, Science, 2006

Biomarkers and surrogate endpoints

- Biomarkers are characteristics that are objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention

- Biomarkers as surrogate endpoints are biomarkers that are intended to substitute for clinical endpoints. Surrogate endpoints are expected to predict clinical benefit (or harm or lack of benefit or harm) based on epidemiologic, therapeutic, pathophysiologic, or other scientific evidence.

Prentice's criteria

- For a given treatment (Z), a surrogate (S) may be validly substituted for a true endpoint (T) if and only if:
  1. *P(S|Z) ≠ P(S)* ... *Z has significant effect on S*
  2. *P(T|Z) ≠ P(T)* ... *Z has significant effect on T*
  3. *P(T|S) ≠ P(T)* ... *S has significant effect on T*
  4. *P(T|S,Z) = P(T|S)* ... *Z has no effect on T given S*

- Ideal biomarkers - entire treatment effect Z on true endpoint T is captured by surrogate S (100% explained)
  - A valid surrogate is defined as a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint

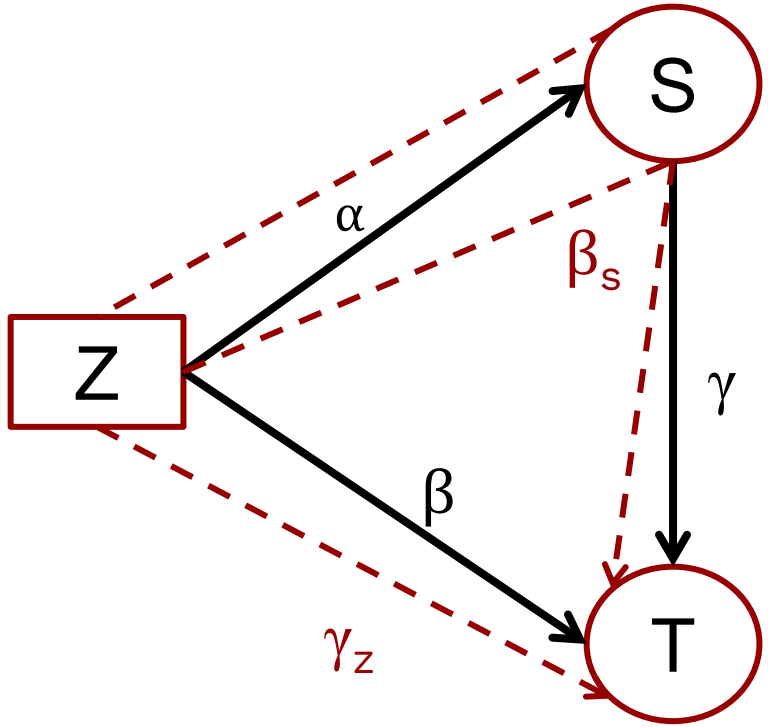
Prentice 1989, Stat Med 8(4): 431.

Real world biomarkers

- Z: Treatment
- S: Surrogate Endpoint
- T: True Endpoint

- α: Effect of Z on S
- γ: Effect of S on T
- β: Effect of Z on T

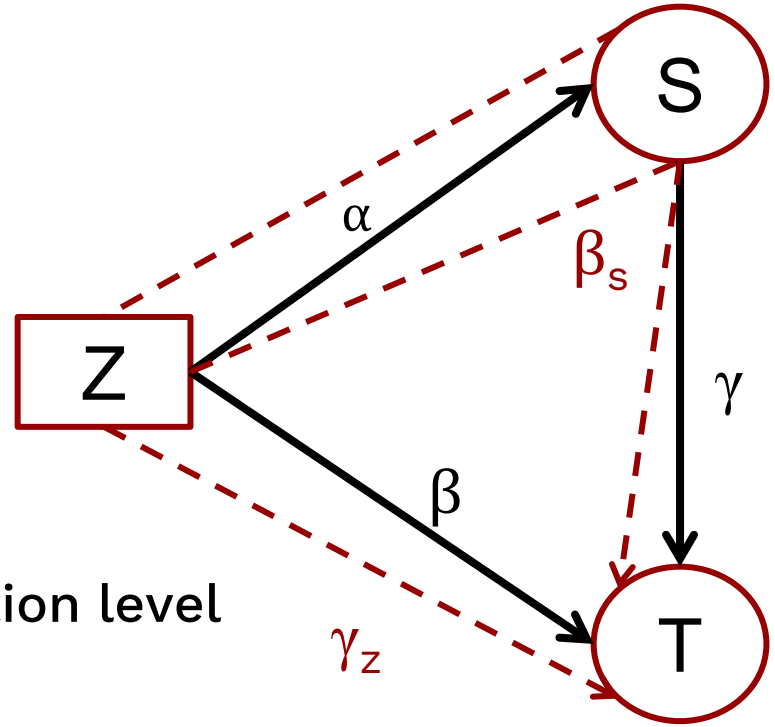
- β_S: Effect of Z on T Regardless of S
- γ_Z: Effect of S on T Regardless of Z



Buyse and Molenberghs 1998, Biometrics 54(3):1014

Real world biomarkers

- **Proportion explained**
  - $PE = \frac{\beta - \beta_S}{\beta} = 1 - \frac{\beta_S}{\beta}$
  - PE → 1 when $\beta >> \beta_S$
  - (Prentice requires $\beta_S = 0$)
- **Relative effectiveness**
  - $RE = \frac{\beta}{\alpha}$
  - RE = 1 → perfect surrogate a population level
- **Adjusted association**
  - AA = γ_Z
  - γ_Z → infinity = perfect surrogate at individual level

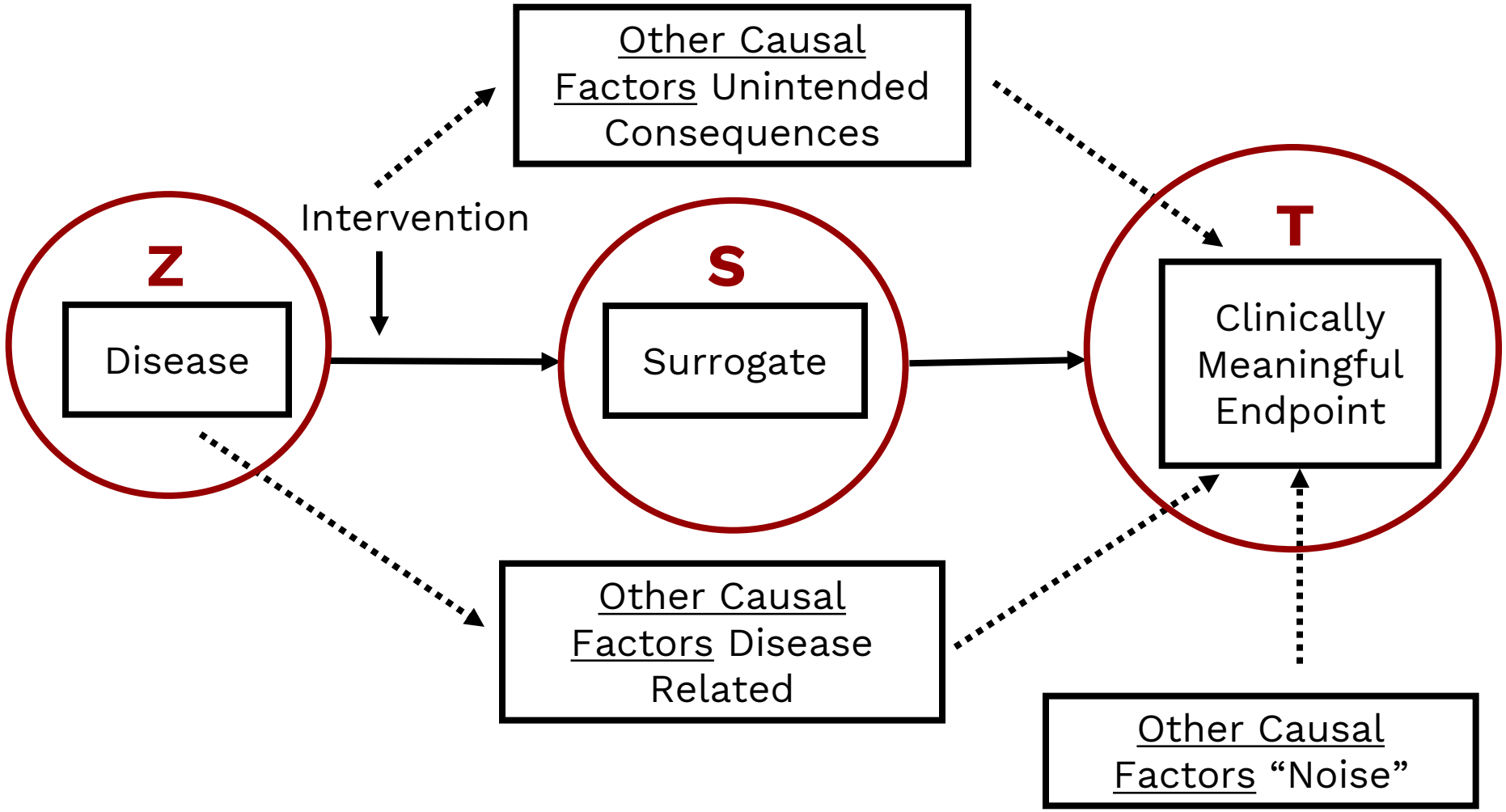


Buyse and Molenberghs 1998, Biometrics 54(3):1014

Types of biomarkers

- Risk assessment biomarkers leading to preventive interventions for those at sufficient risk
- Early detection biomarkers enabling intervention at an earlier and potentially more curable stage than under usual clinical diagnostic conditions
- Prognostic biomarkers allowing for more aggressive therapy for patients with poorer prognosis
- Predictive biomarkers of response to a therapy, thereby providing guidance in choice of therapy
- Treatment response biomarkers monitoring the response of disease during therapy, with potential for adjusting level of intervention (e.g. dose) on a dynamic and personal basis
- Recurrence biomarkers enabling early detection of recurrence

Ideal vs. Real world biomarkers



School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MADISON

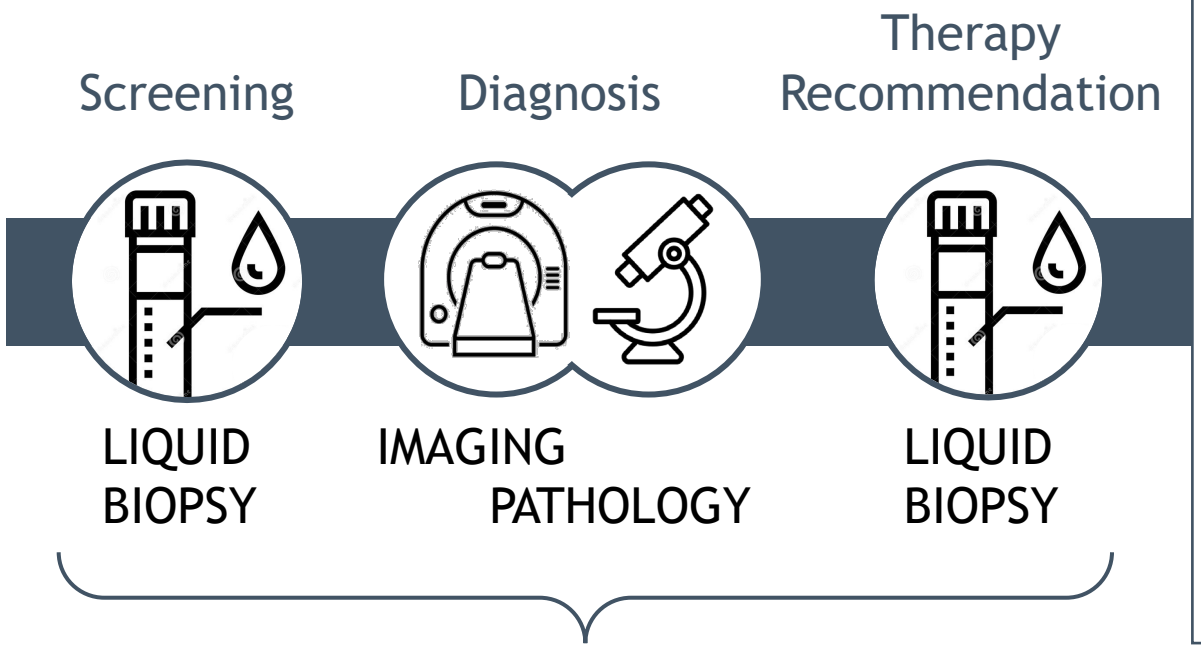
Prentice 1989, Stat Med 8(4): 431.

Biomarker characteristics

- **Clinical relevance**
  - Firm biological rationale
  - Need to get information quickly
- **Sensitivity / specificity to treatment effects**
  - Correlated with outcome data
- **Reliability**
  - Measured with accuracy, precision, reproducibility
  - Uncertainties understood and quantified
- **Practicality / simplicity**
  - Tolerated by patients
  - Easily integrated into clinical workflow

Lesko and Atkinson 2001, Annu Rev Pharmacol Toxicol 41:347.

What biomarkers can we use?



Modalities ideally suited for early in patient journey

LIQUID BIOPSY

- Early detection
- Initial therapy selection (genomics)

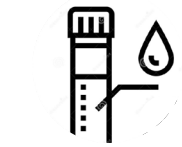
PATHOLOGY

- Diagnose/confirm disease
- Microtumor environment
- Initial therapy selection

IMAGING

- Detect disease/lesions
  - Manual (standard of care)
  - Artificial intelligence systems

What biomarkers can we use?



LIQUID BIOPSY

- 👍 Low cost/high frequency
- 👎 Single response metric for entire patient
- 👎 No spatial information
- 👎 Limited understanding of heterogeneity

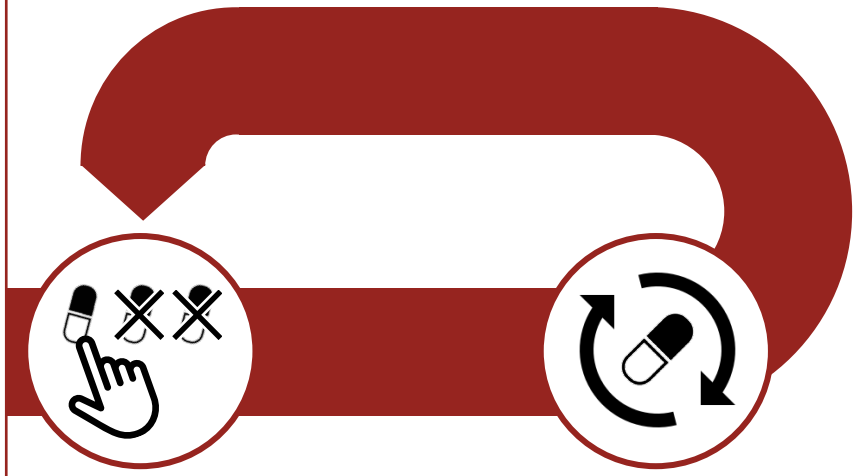
PATHOLOGY

- 👍 Analysis of disease evolution
- 👎 Information limited to lesions sampled
- 👎 Invasive

IMAGING

- 👎 Assessment limited to 3-5 lesions
- 👎 Poor predictive accuracy

} Manual systems



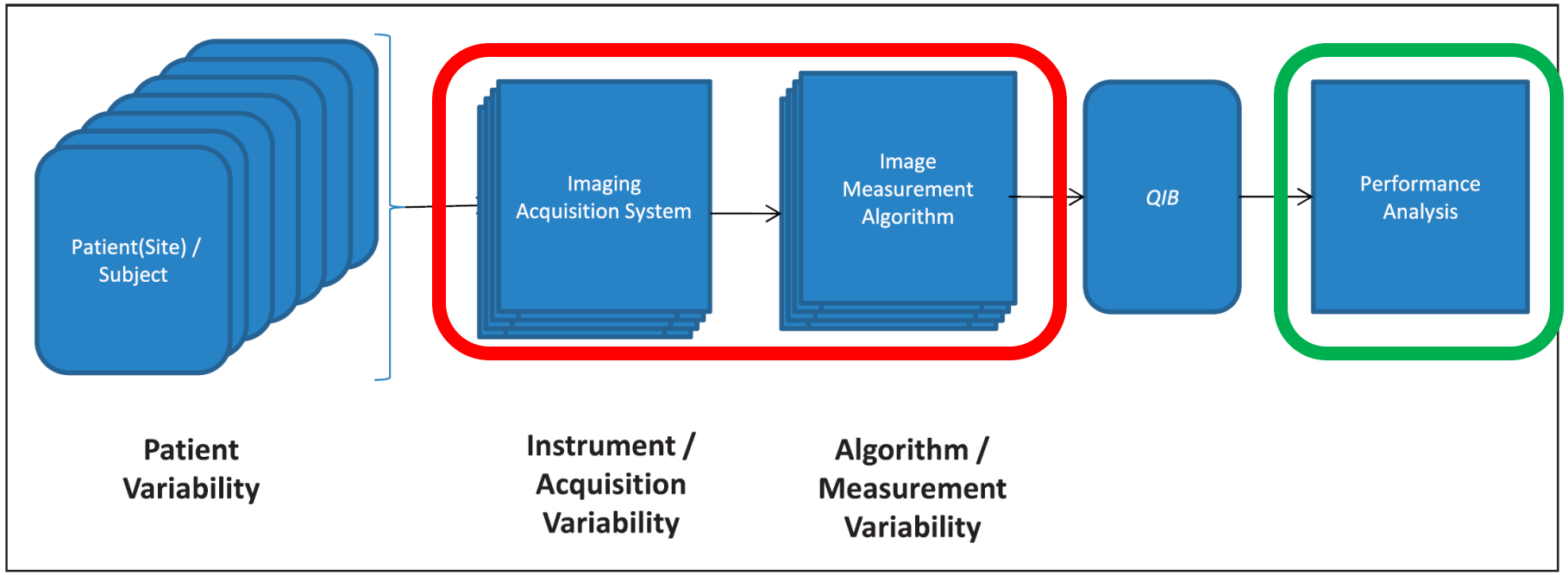
Therapy Selection

Therapy Optimization

QUANTITATIVE IMAGING BIOMARKERS

- Biomarkers vs surrogate endpoints
- Types of biomarkers

Quantitative Imaging Biomarkers (QIB)



Raunig et al 2015, Stat Methods Med Res 24(1): 27-67

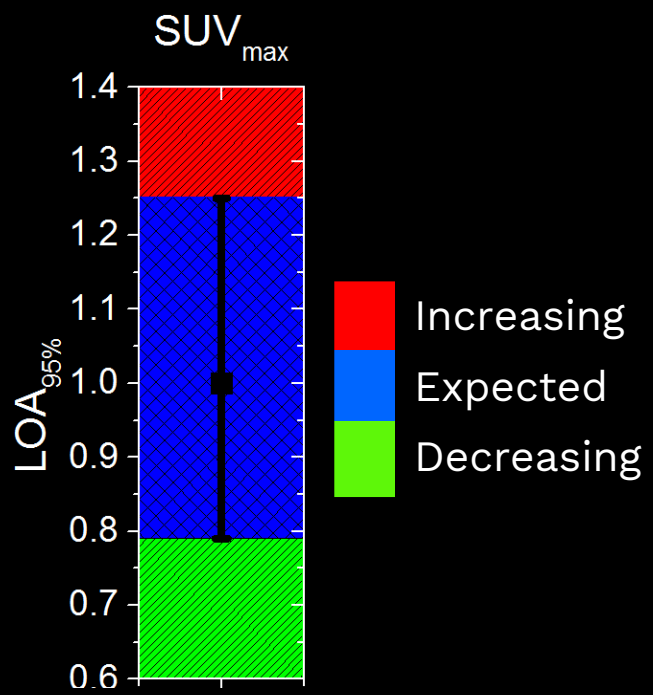
QIB uncertainties –> significant changes

Limits of agreement:

$$LOA_{95\%} = (e^{(B-RC)}, e^{(B+RC)})$$

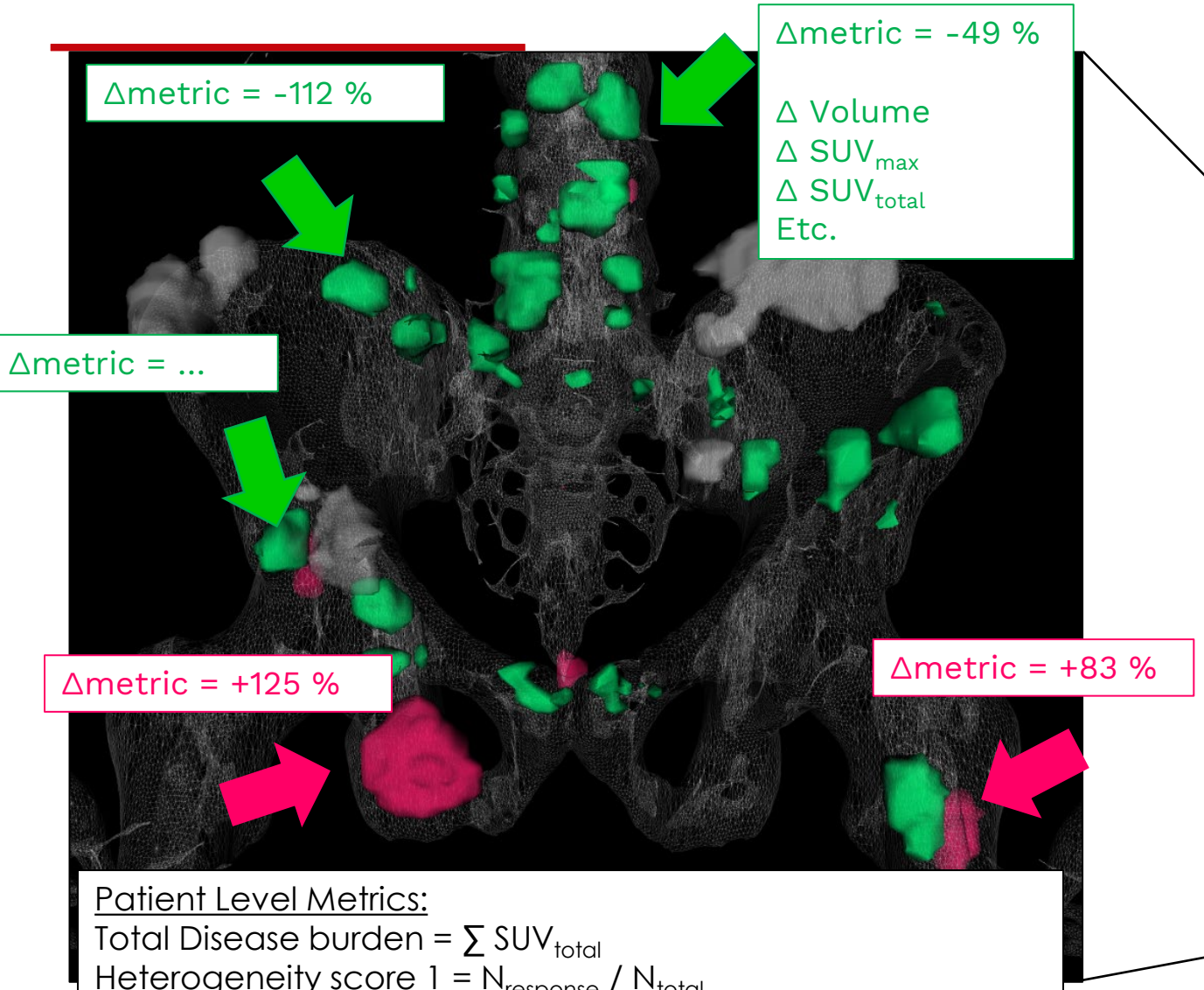
$$B = \frac{m_2}{m_1}$$

$$RC = 1.96\sqrt{\sigma}$$



Test-retest **limits of agreement (LOA)** are used to define **significant changes** in imaging features

QIBs allow quantification of response



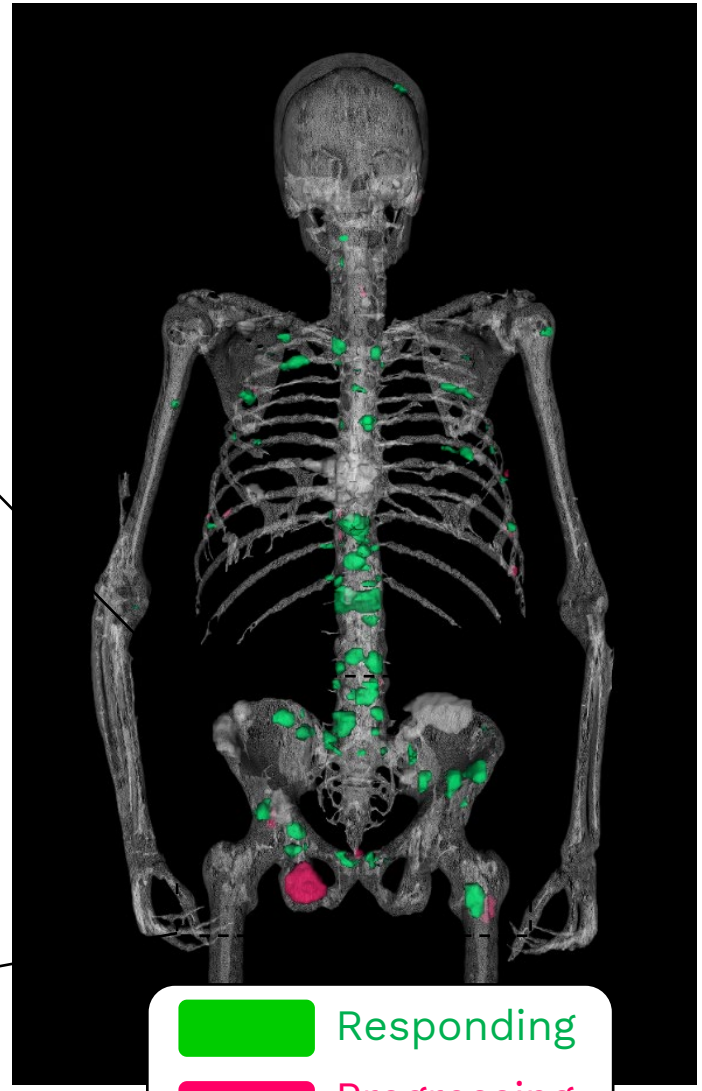
Patient Level Metrics:

Total Disease burden = $\sum SUV_{total}$

Heterogeneity score 1 = $N_{response} / N_{total}$

Heterogeneity score 2 = $N_{resistant} / N_{total}$

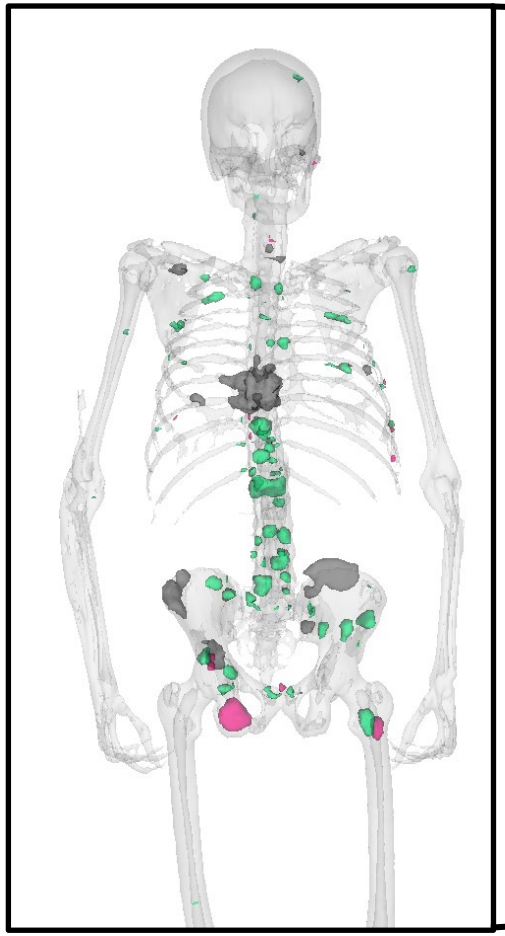
Response Map

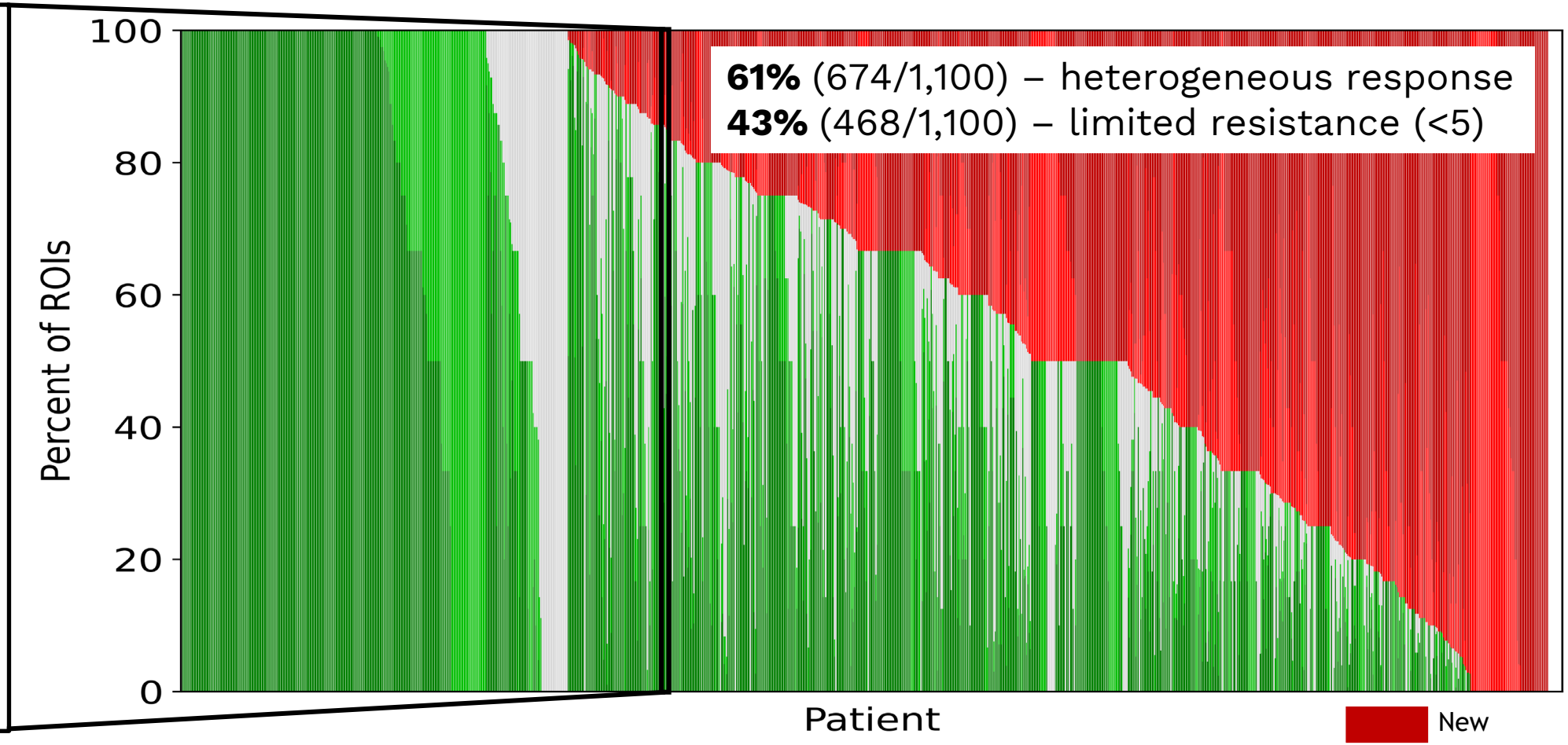


THE ISSUE WITH QIB IN PRECISION MEDICINE

- AI-based Treatment Response Assessment

QIB for precision medicine:
Inter-tumor response heterogeneity





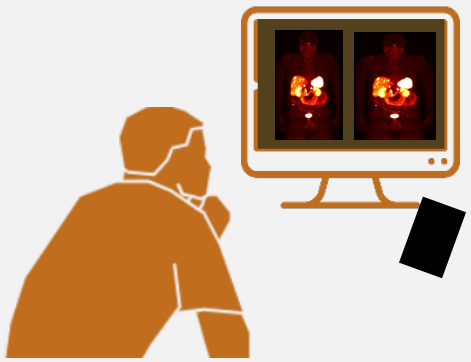
Cancers included (n=1,100): prostate, bladder, renal, testicular, penile H&N, NSCLC, NET, melanoma, ACC

Miyahira and Soule 2024, Prostate, 1-19

School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MADISON

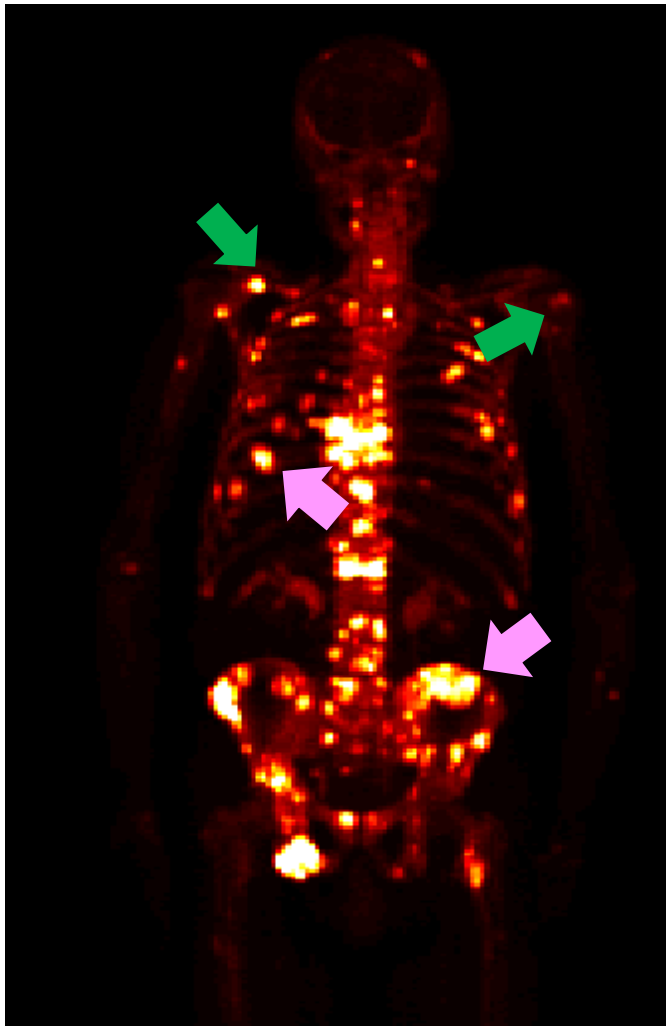
Treatment response assessment –
Current practice

Manual and Qualitative Assessment

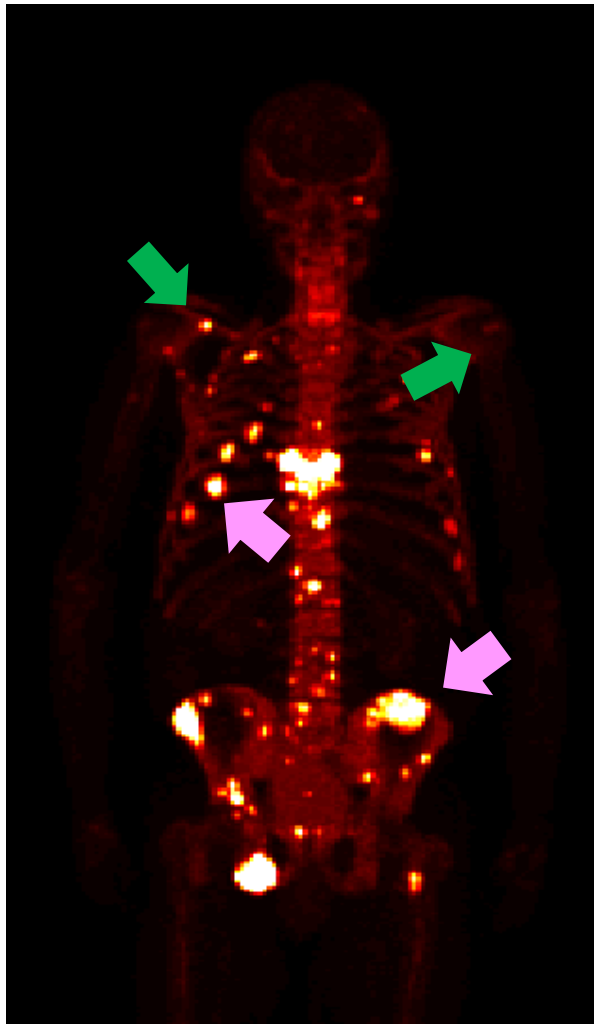


Radiologists/nuc med physicians manually identify subset of lesions for treatment evaluation

Time point 1



Time point 2



What information do we want to extract from imaging data?

Number of lesions?

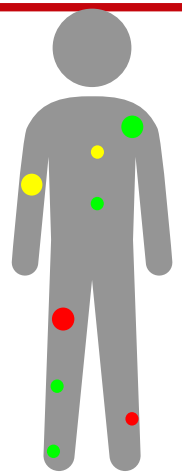
Total disease burden?

Inter-lesion heterogeneity?

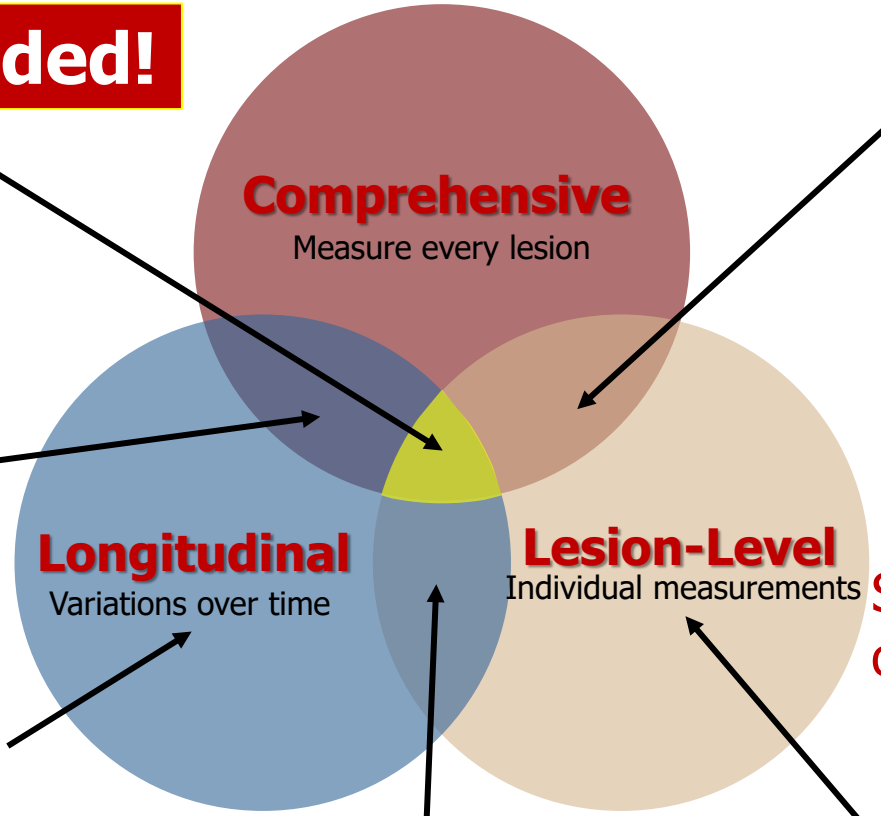
……

How can we capture this information efficiently and objectively?

Treatment response assessment –
State-of-the-art



This is needed!



Segment all lesions on baseline
Ohlendorf, F. (2020) QJNMMI
Carlsen, E. (2021) J Nuc. Med.

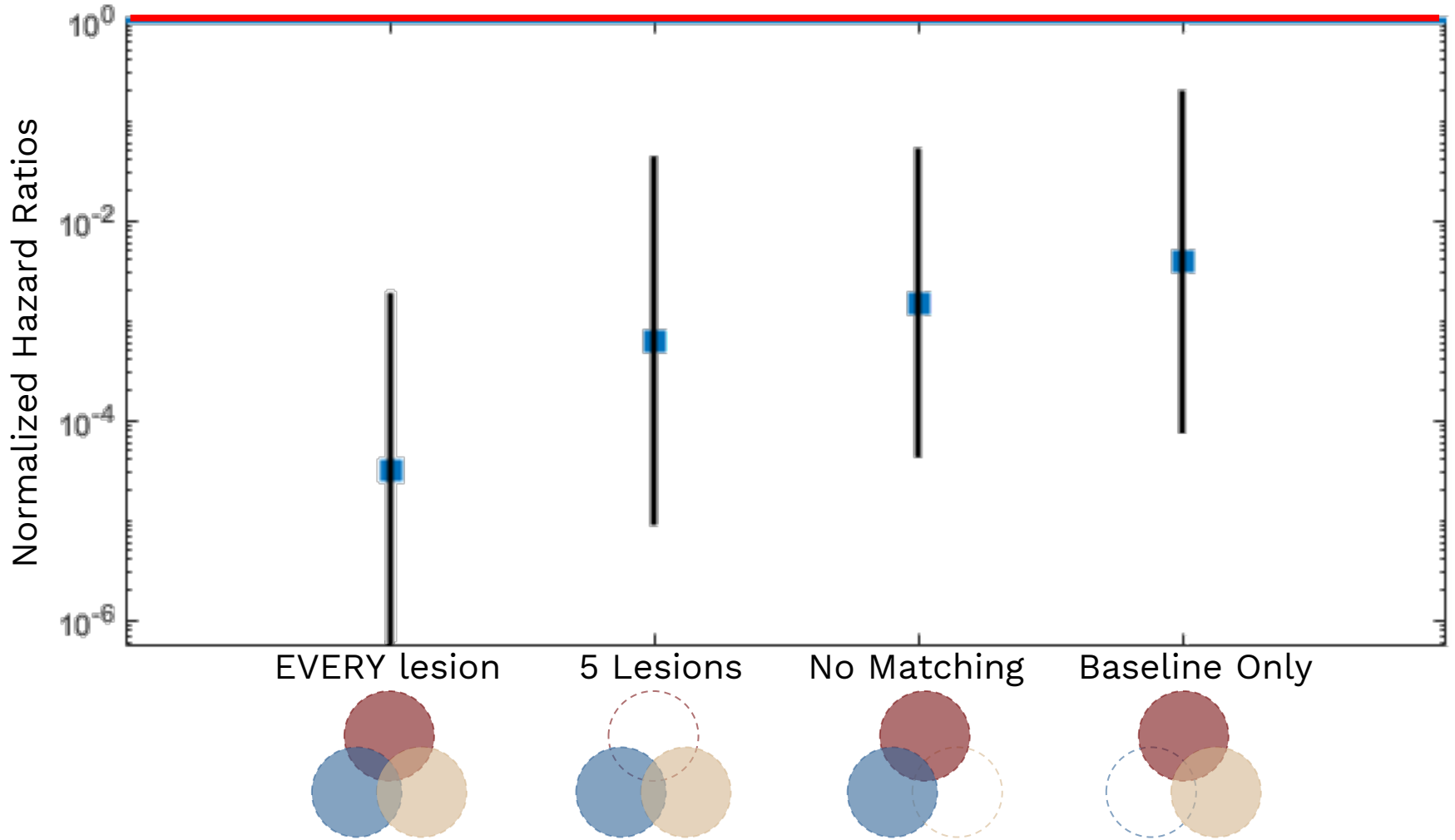
Segment all lesions
Batched longitudinal metrics
Pauwels, E. (2020) J Nuc. Med.

Segment some lesions
on baseline images
Campana, D. (2010) J Nuc. Med.
Ambrosini, V. (2015) J Nuc. Med.
Werner, R. (2017) Oncotarget
Werner, R. (2019) Molec. Imag. And Biol.
Graf, J. (2020) Eur. J Nuc. Med. Molec. Imag.
Zwirtz, K. (2022) Pharmaceutics

Segment 5 lesions
Batched longitudinal metrics
Kratochwil, C. (2015) Molec. Imag. and Biol.
Sharma, R. (2019) RT and Onc.
Ortega, C. (2021) J Nuc. Med.
Urso, L. (2023) Diagnostics

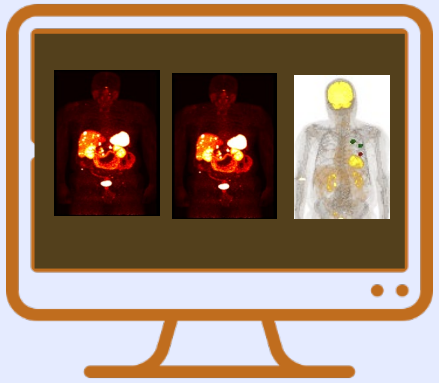
One lesion!
Gabriel, M. (2009) J Nuc. Med.
Haug, A. (2010) J Nuc. Med.

Why we need to assess EVERY lesion?



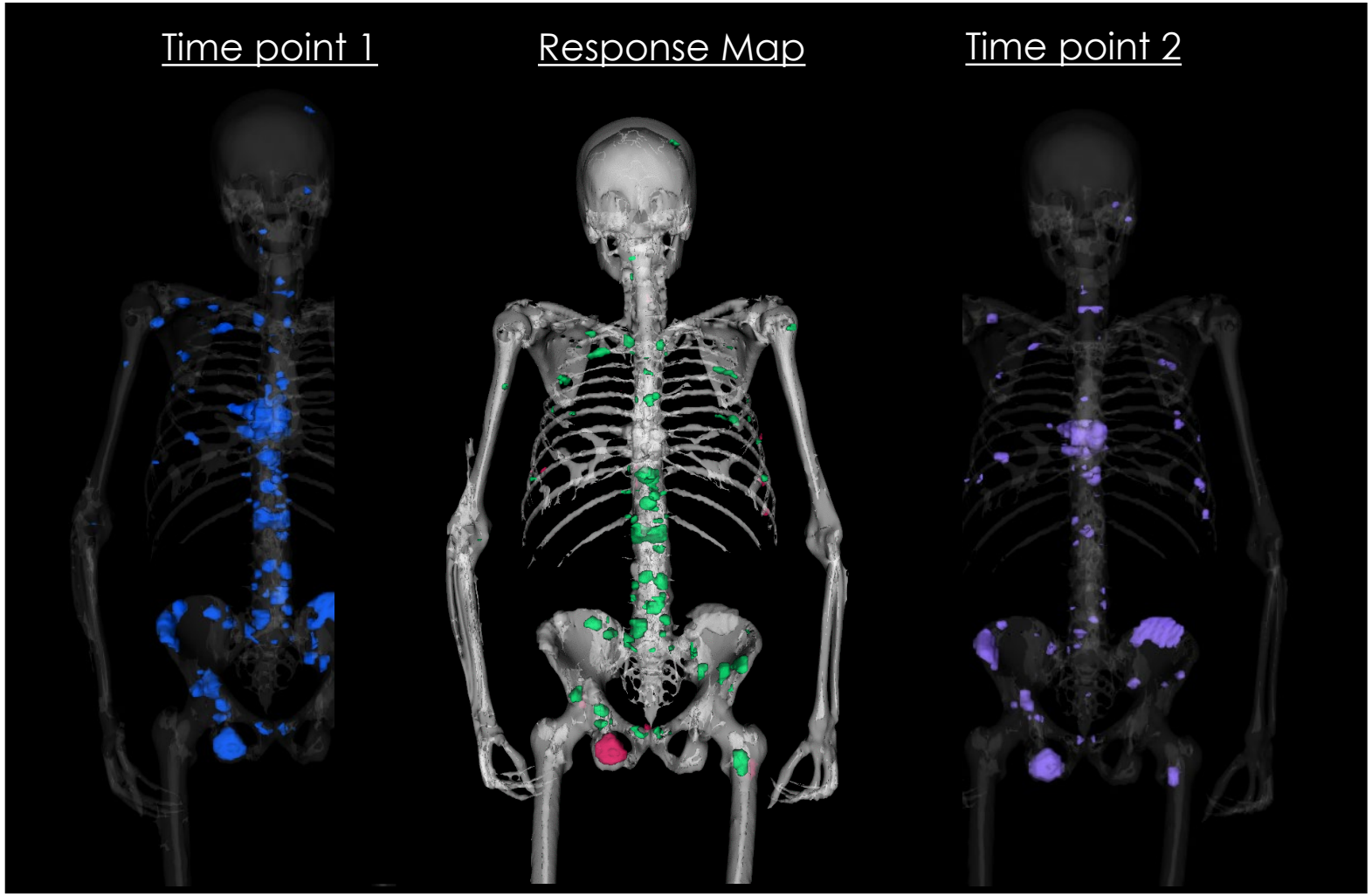
Treatment response assessment – AI-based approach

Automatic and Quantitative Assessment



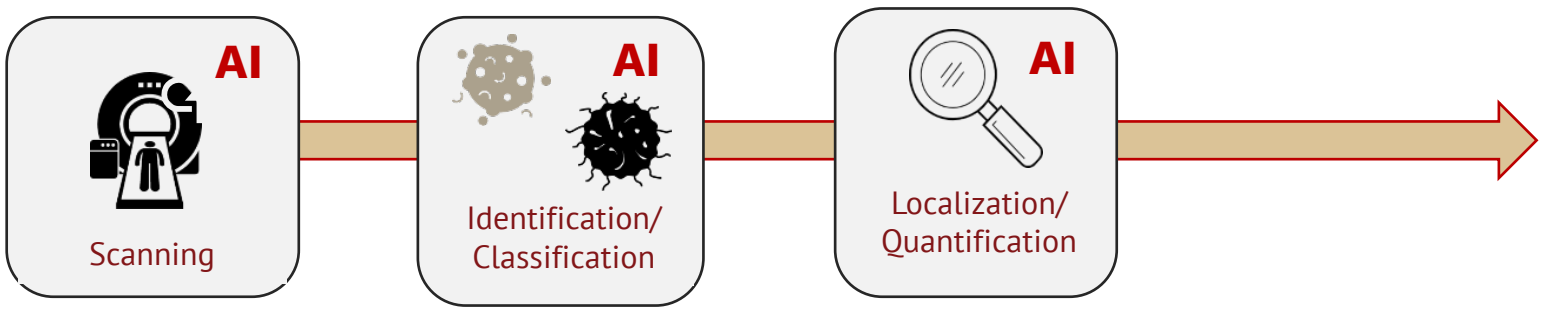
Our software automatically detects and classifies all lesions

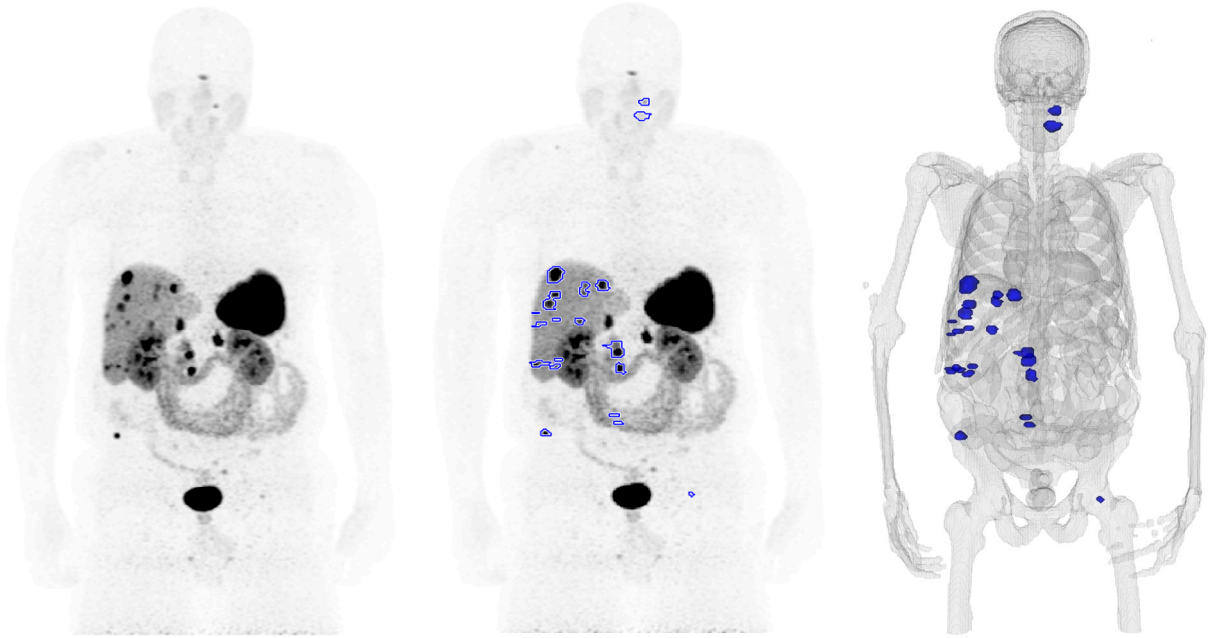
US Patents 9603567, 10445878
Licensed to our spin-off:
AIQ Solutions

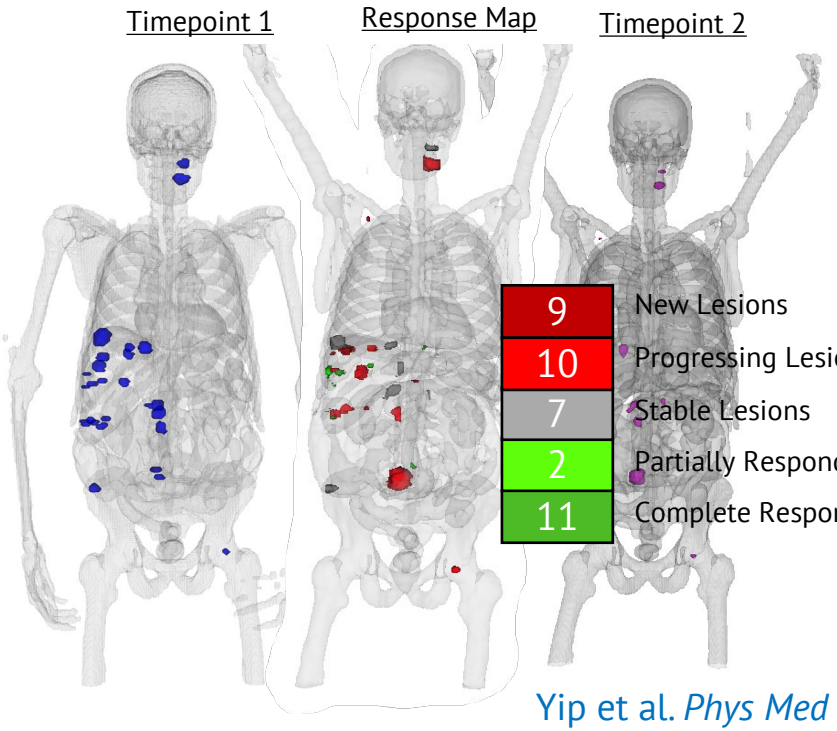


School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MADISON

Treatment response assessment – AI-based workflow

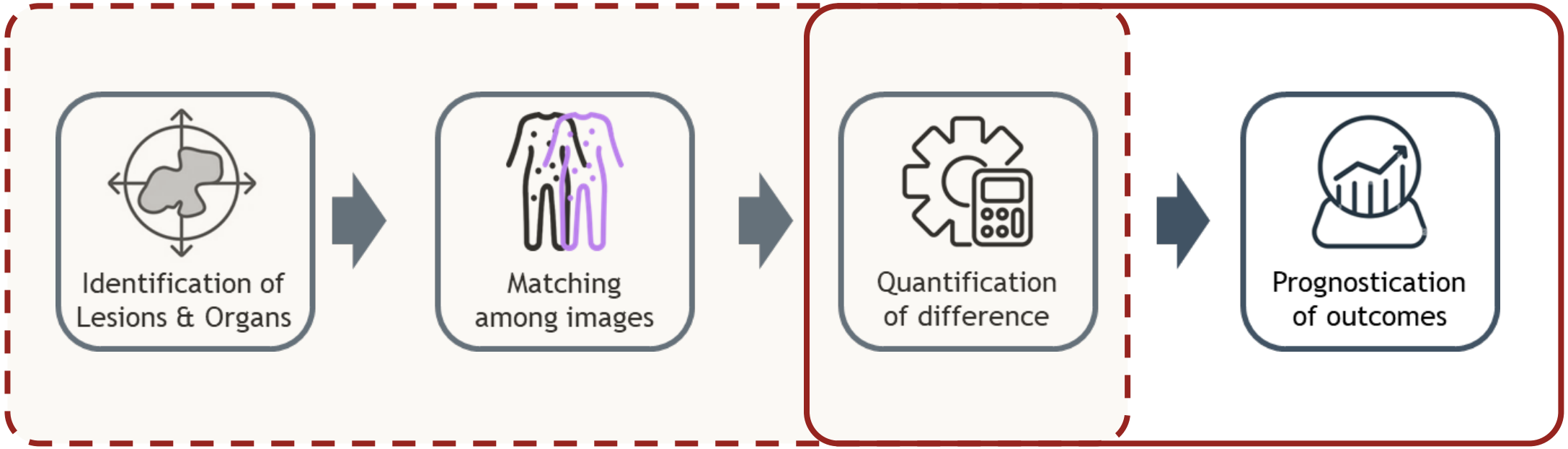






Yip et al. *Phys Med Biol*. 2014
Santoro-Fernandes et al. *Phys Med Biol*. 2021
US Patent 9,603,567

AI-driven association with outcomes

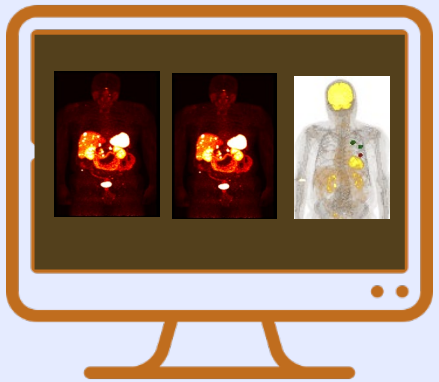


Quantitative Imaging Biomarkers

Surrogate Endpoints
(Predictive Biomarkers)

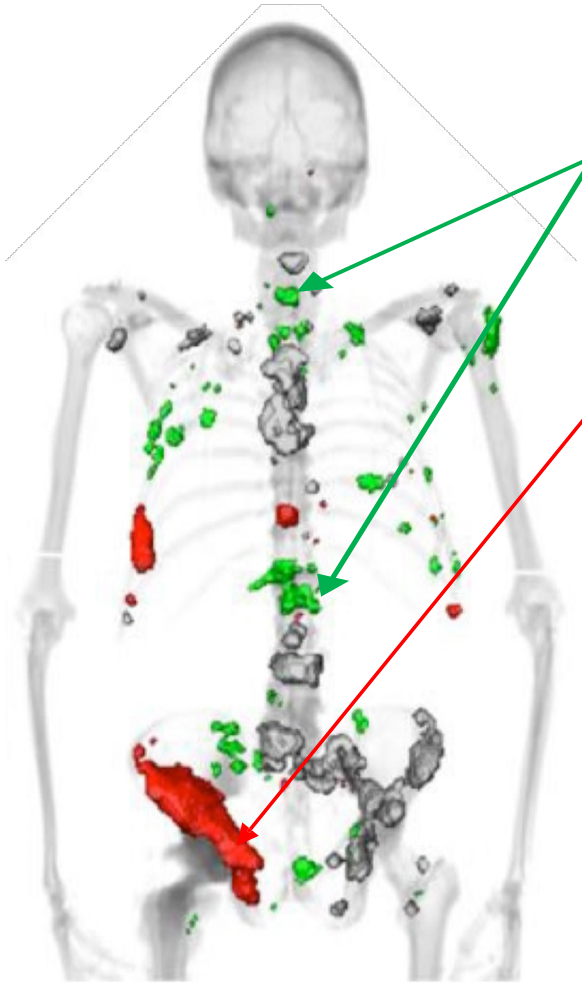
AI-based Quantitative Imaging Biomarkers

Automatic and Quantitative Assessment



Our software automatically detects and classifies all lesions

US Patents 9603567, 10445878
Licensed to our spin-off:
AIQ Solutions

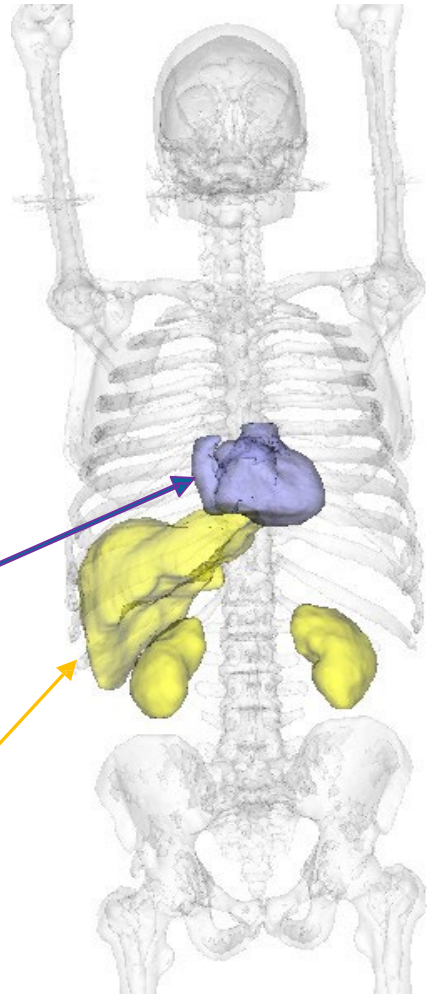


Responding to treatment

Resistant to treatment

Medium toxicity risk

High toxicity risk



School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MADISON

HOW CAN WE SAFELY DEPLOY AI-BASED QIB?

- Out Of Distribution (OOD) uncertainty
- In Distribtuion (ID) uncertainty

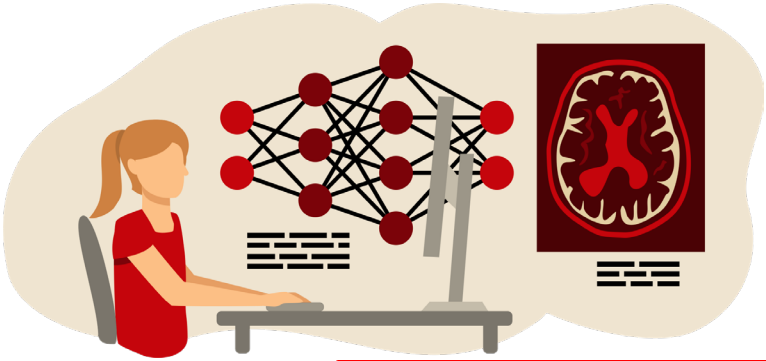
The critical component of AI-QIBs...

What do the people want? — **How are we going to get there?** — **Is that it? Are we there?**

Clinical DL Deployment

https://www.gep.com/blog/mind/artificial-intelligence-step-forward-clinical-trials

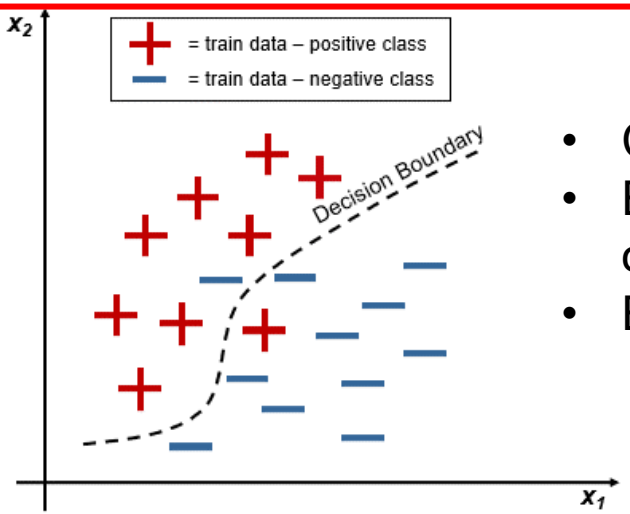
Current DL Models



Proceed Forward

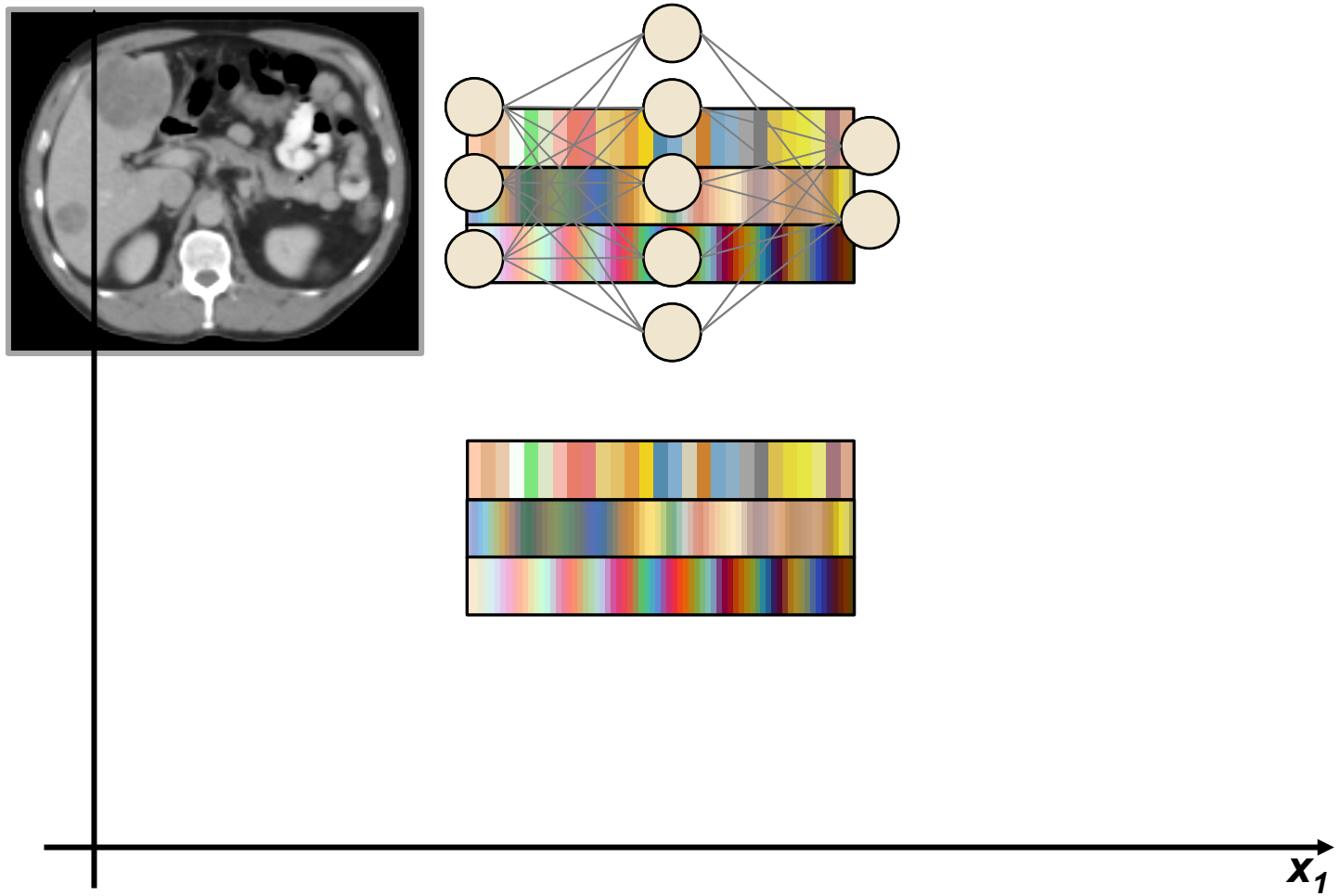
AI Model Uncertainty Quantification

What's missing?



- Catch "silent errors"
- Enhance user trust in output
- Enhanced interpretability

AI-QIB uncertainties and data domain

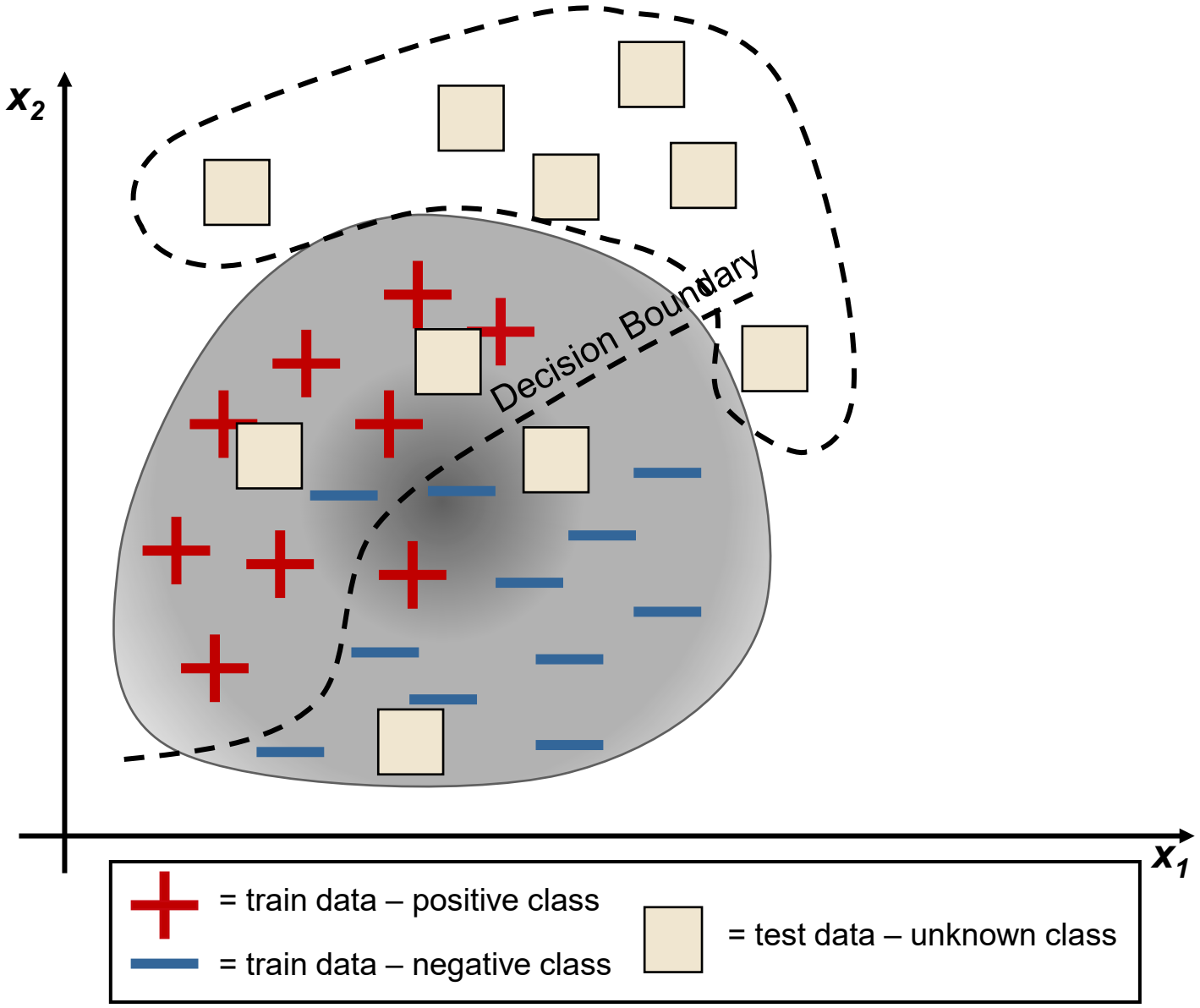


AI-QIB uncertainties and data domain

Data Domain:
- Set of data features present in a train dataset

Out-of-Domain (OOD):
- **Data:** Test samples with *dissimilar* features as the train data
- **Uncertainty:** Due to model unfamiliarity

In-Domain (ID):
- **Data:** Test samples with *similar* features as the train data
- **Uncertainty:** Model fitting limits, data noise

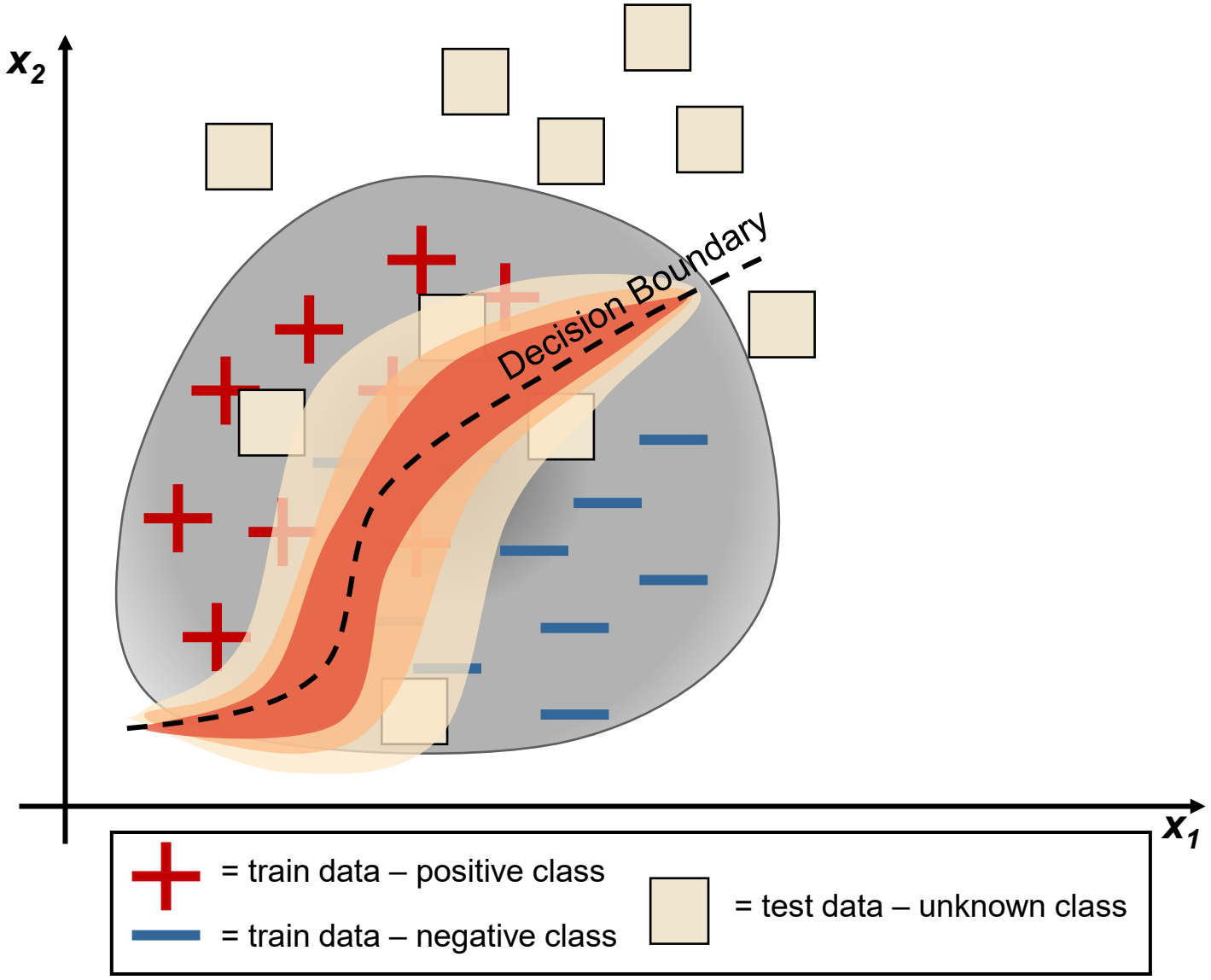


AI-QIB uncertainties and data domain

Data Domain:
- Set of data features present in a train dataset

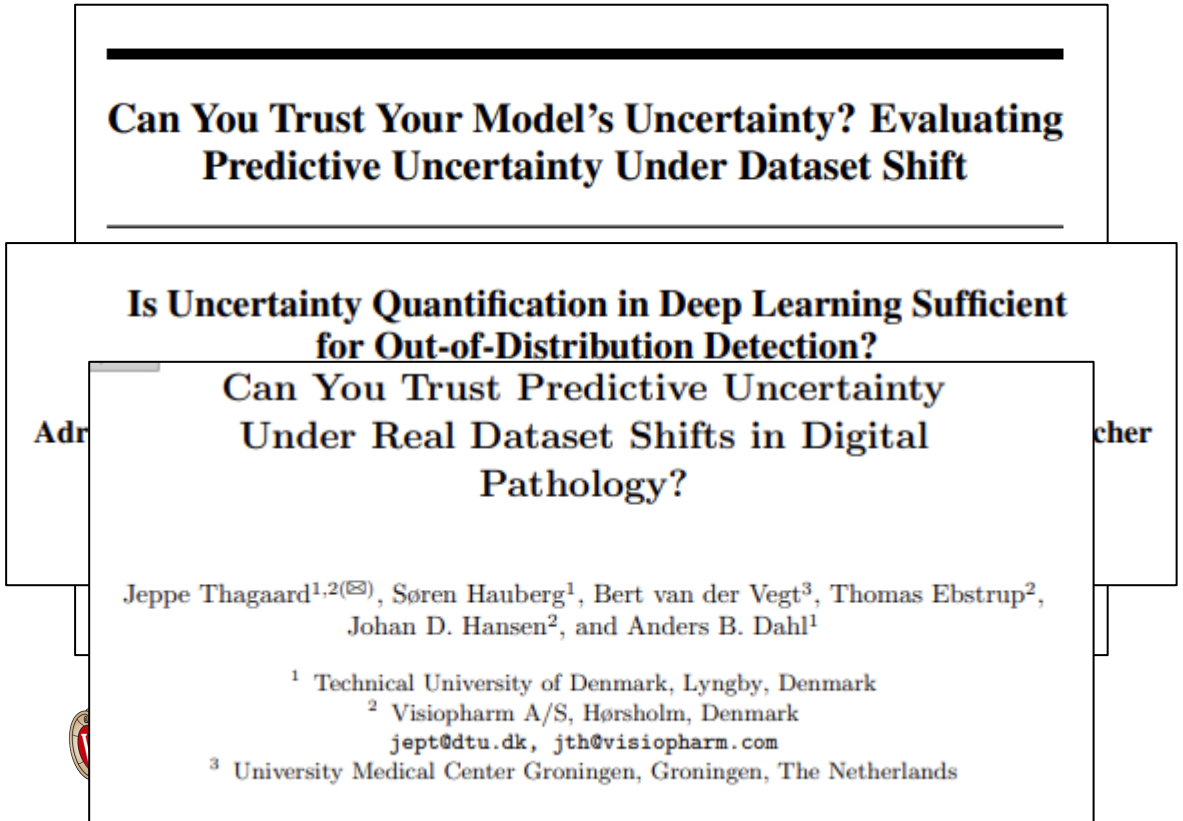
Out-of-Domain (OOD):
- **Data:** Test samples with *dissimilar* features as the train data
- **Uncertainty:** Due to model unfamiliarity

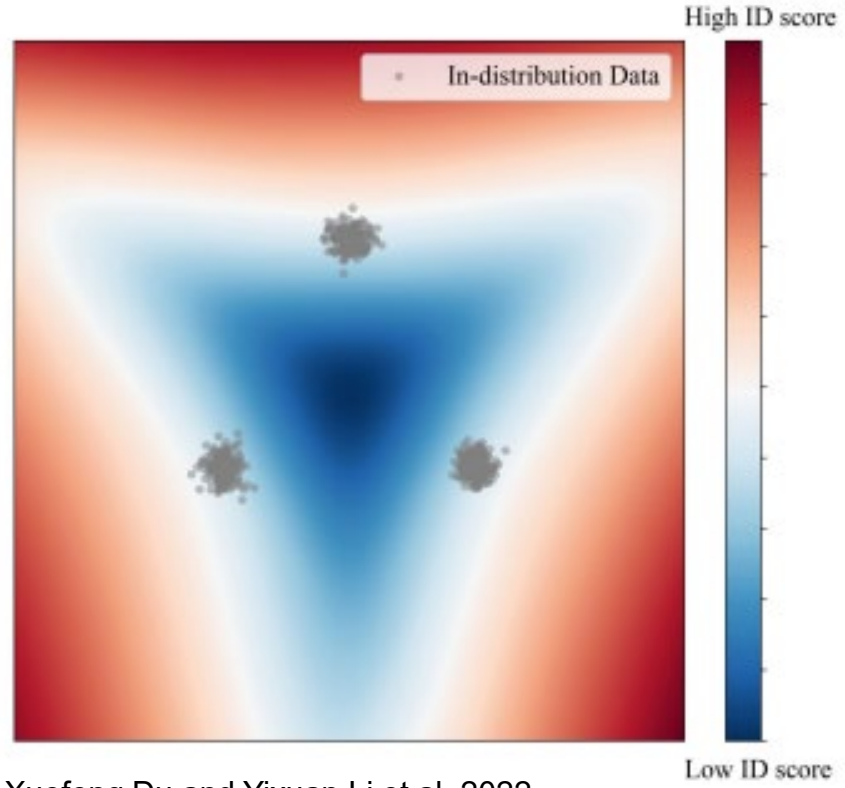
In-Domain (ID):
- **Data:** Test samples with *similar* features as the train data
- **Uncertainty:** Model fitting limits, data noise



Are standard UQ methods enough?

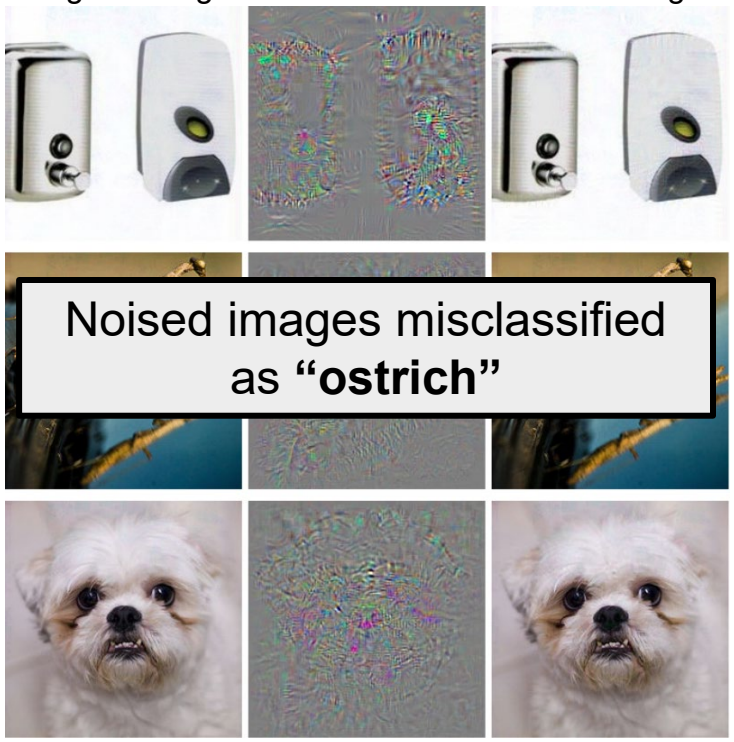
- Most standard UQ methods estimate *predictive uncertainty*
  - E.g., *monte carlo dropout, deep ensembles, test-time augmentation, etc.*
- Work has been done, however, which shows that *predictive uncertainty* methods fail to capture uncertainty due to OOD samples





Xuefeng Du and Yixuan Li et al. 2022

Uncertainties are hard to detect



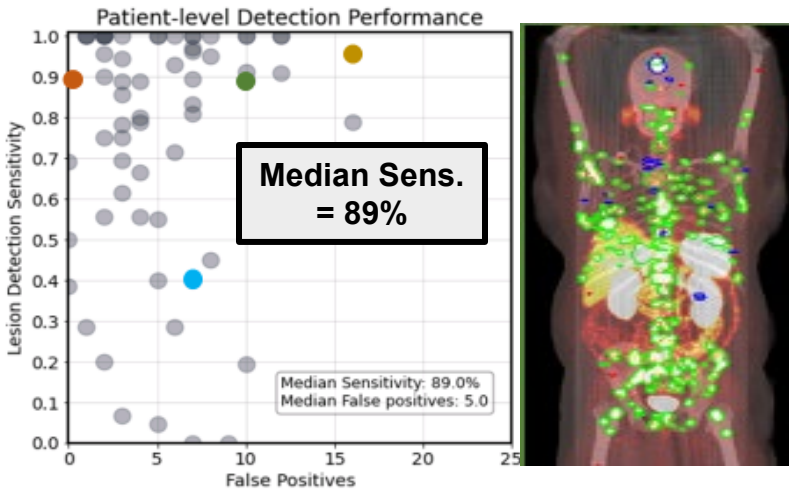
Szegedy, et. al. 2014

Train: *N = 568* ^{68}Ga-DOTATATE PET/CT

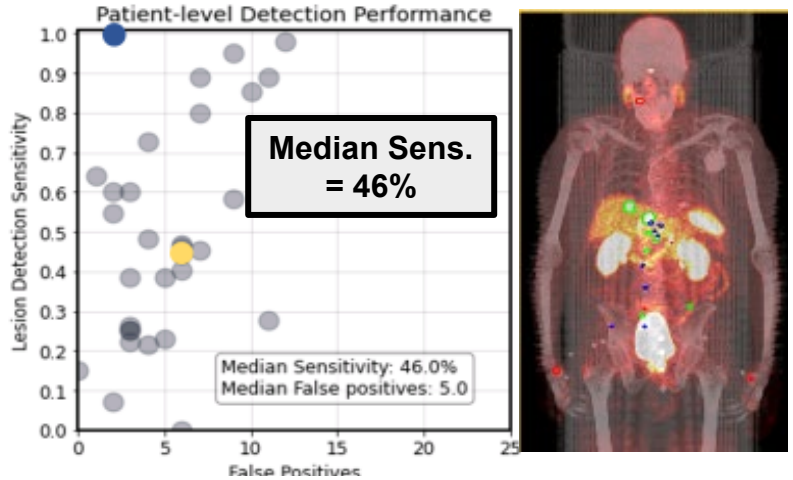
Test 1: *N = 67* ^{68}Ga-DOTATATE PET/CT

Test 2: *N = 31* ^{64}Cu-DOTATATE PET/CT

^{68}Ga-DOTATATE



^{64}Cu-DOTATATE



Clinical OOD Sources

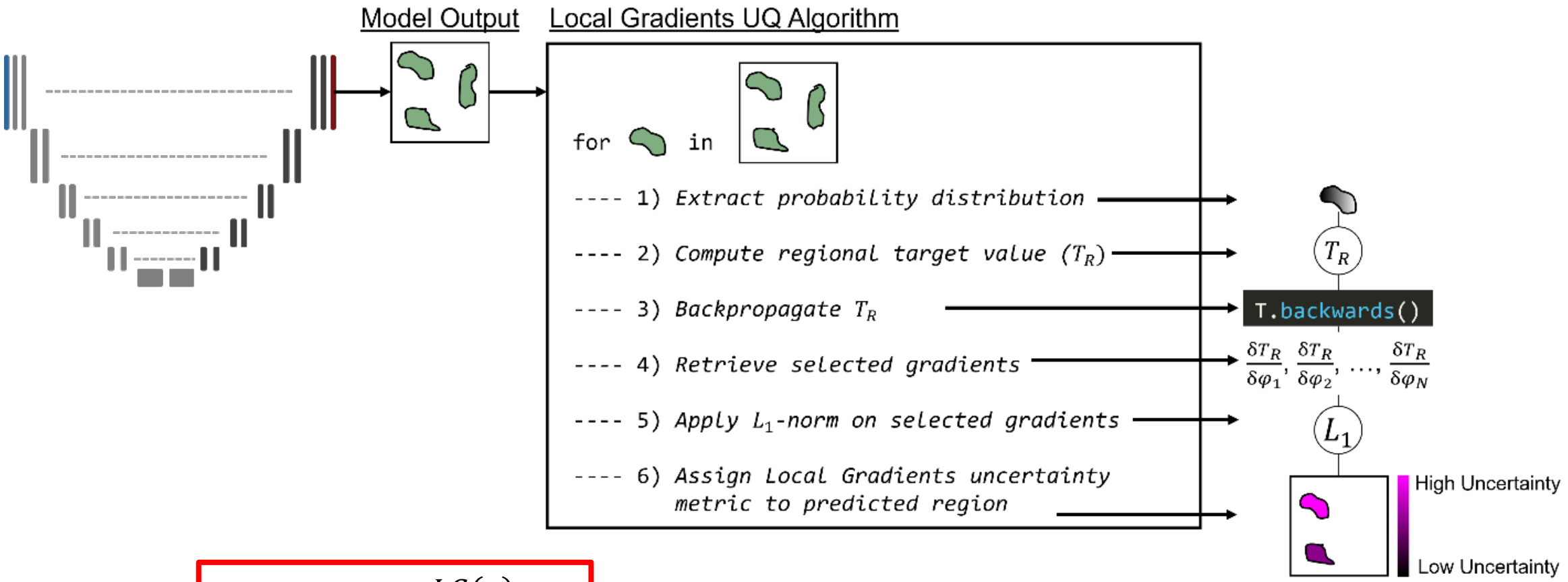
- Small Datasets
- Scanner Differences
- Acquisition Differences

Etc.

Deep learning models fail on OOD data

Medical image models are prone to face OOD data

Local gradients UQ method

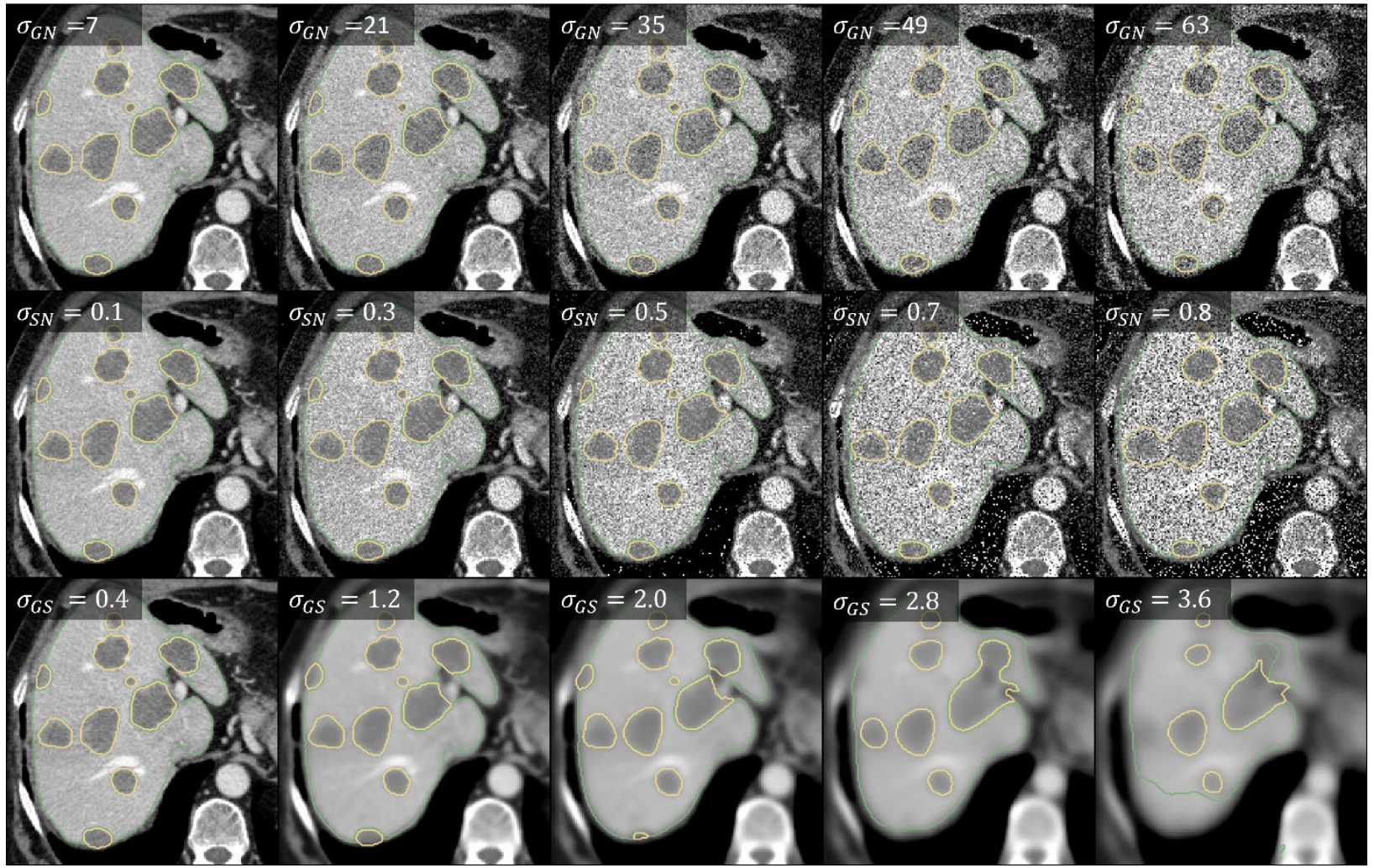


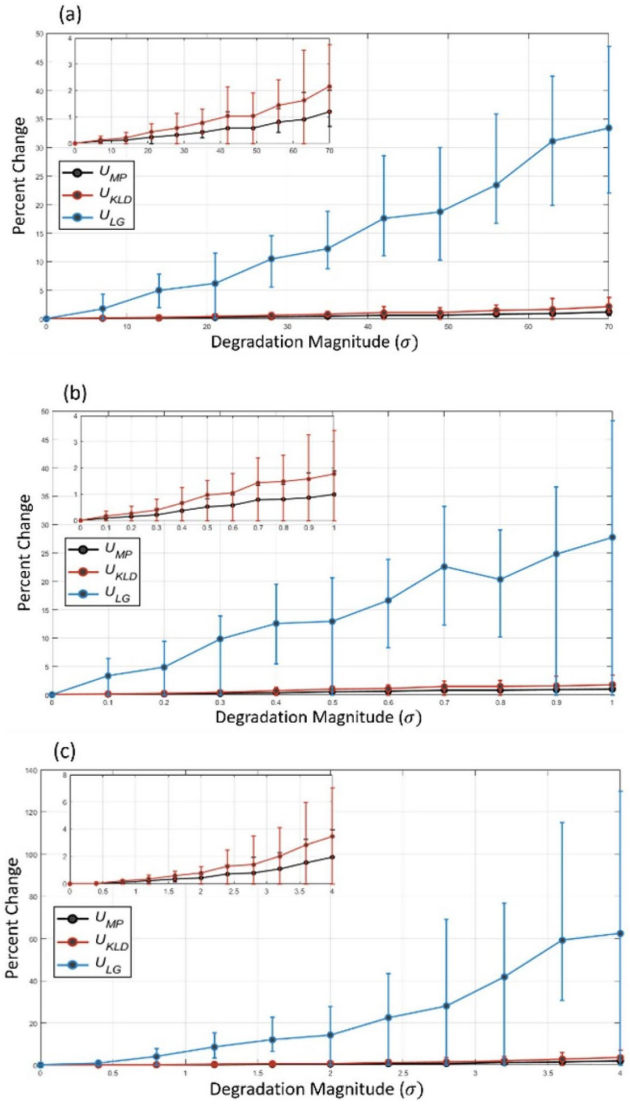
$$U(x)_{LG,R} = \frac{LG(x)_R}{P_{95}\{LG_{low}\}}$$

Local gradients uncertainty measure

Schott et al 2024, Phys Med Biol 69: 155015

Local gradients UQ method

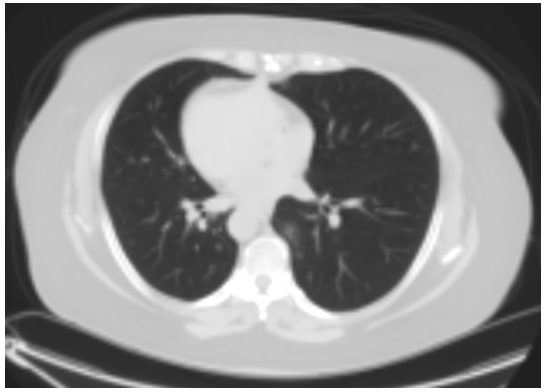




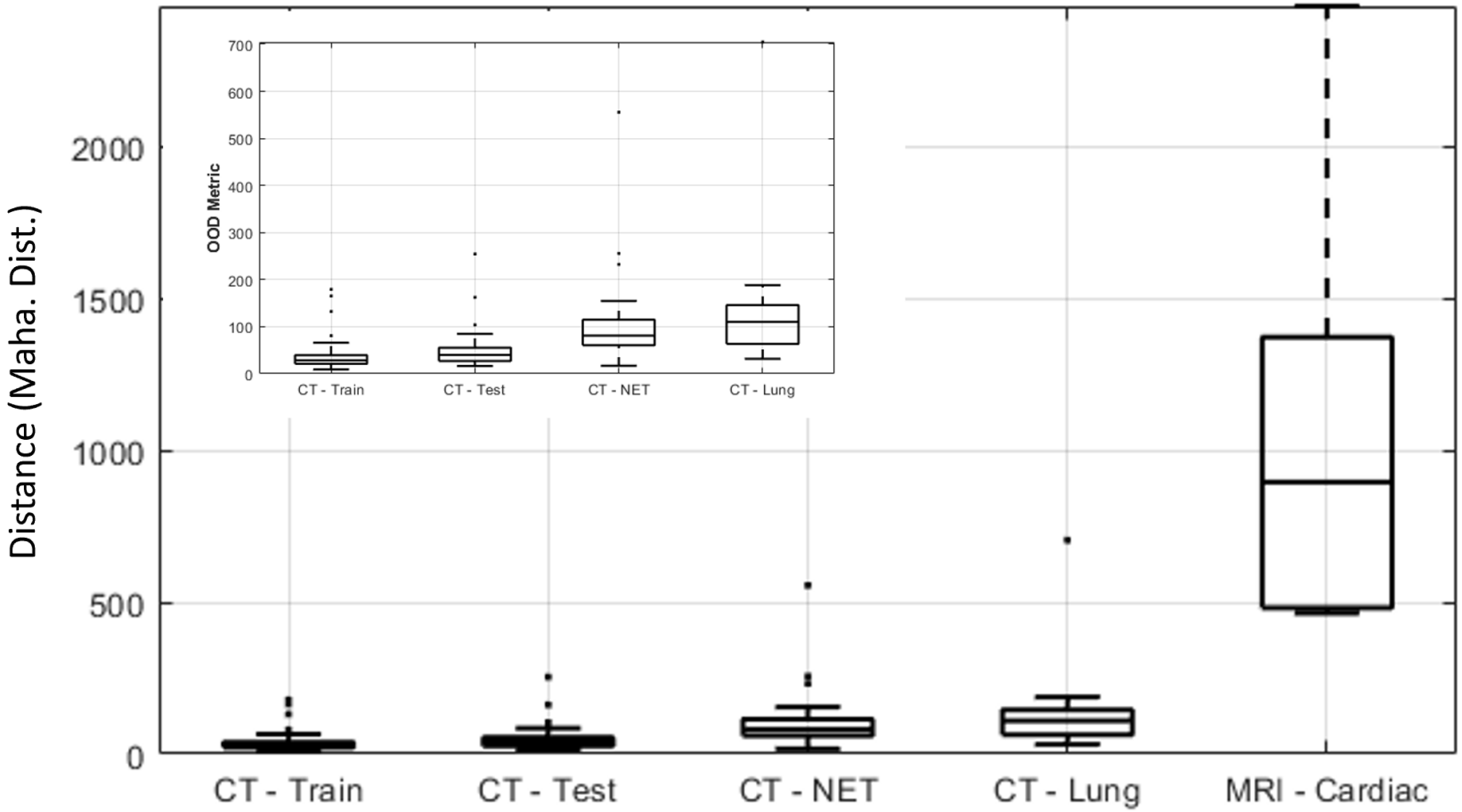
School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MADISON

Schott et al 2024, Phys Med Biol 69: 155015

OOD detection example

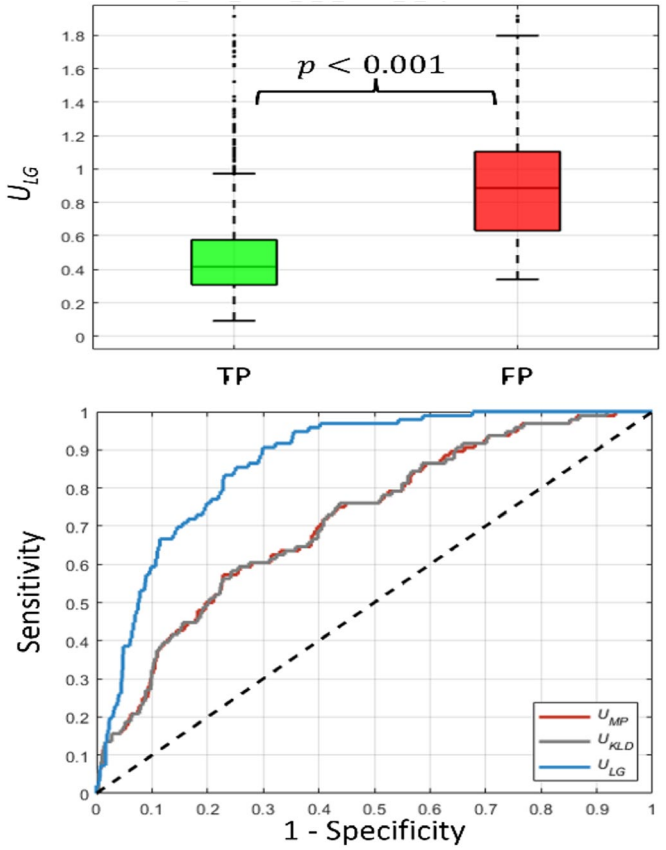
Train: Abdominal CE CT	**Test:** Validation split data	Public CE CT of Liver Tumors	Public Lung CT	Public Cardiac MRI
				
Train data. Defines the ID data distribution	Validation data split off from the train. Likely ID or very near OOD	External data of same disease/modality. Possibly ID→near OOD→far OOD	External data of different anatomy. Likely far OOD	External data of different disease/modality. Likely very far OOD

OOD detection example

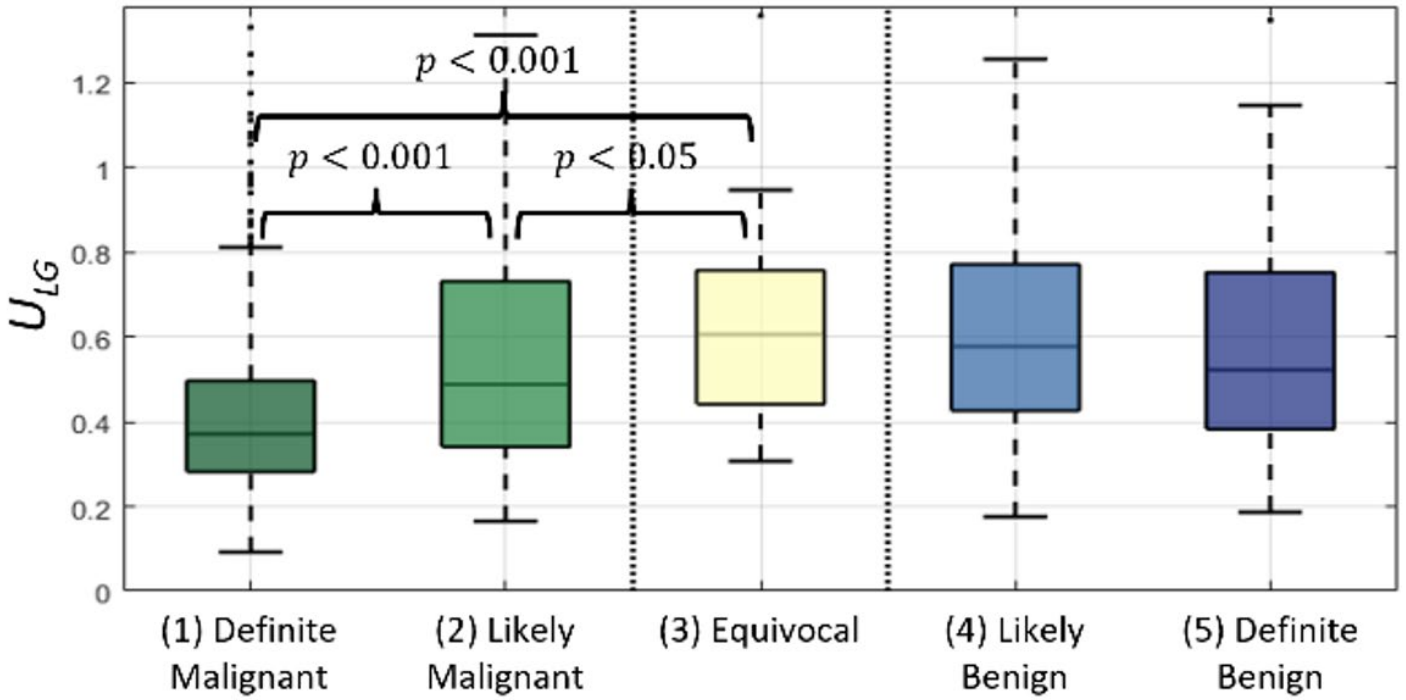


ID uncertainty examples

Identifies False Positives



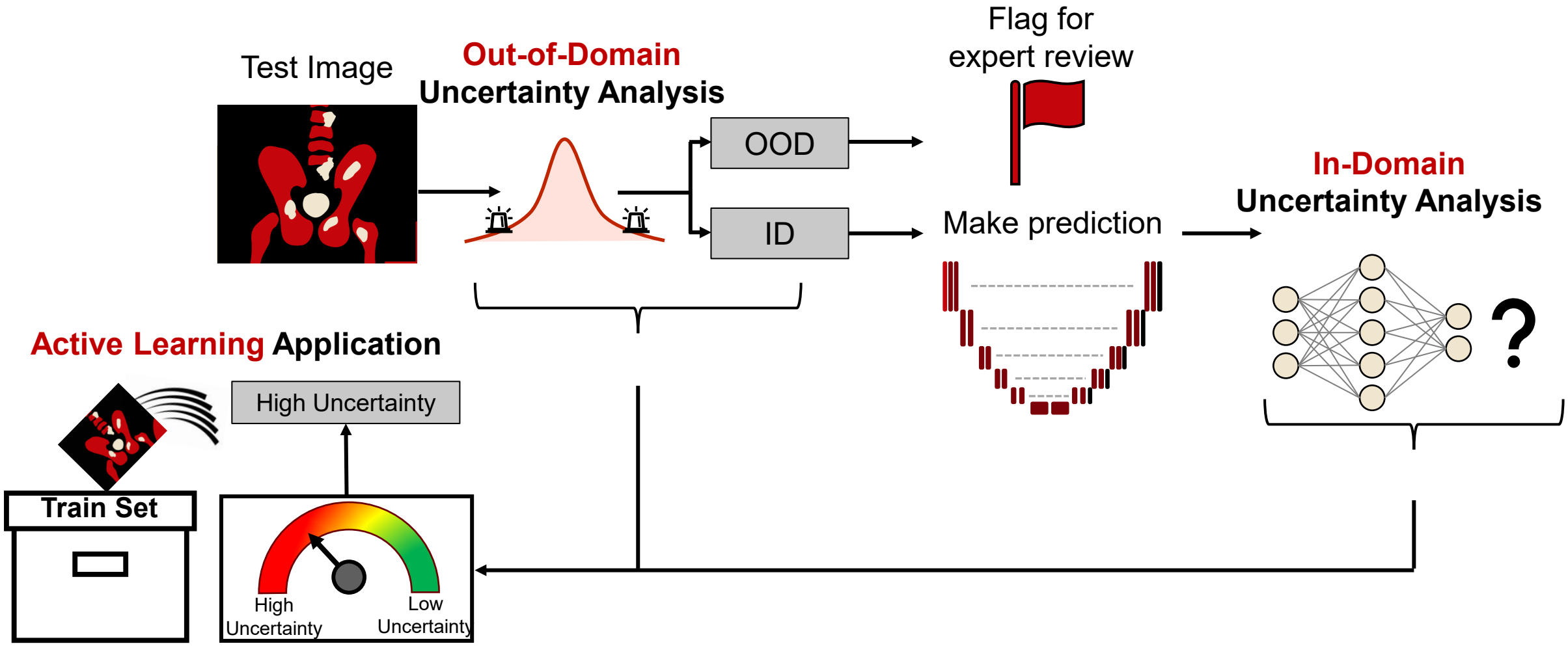
Aligns with physician uncertainty



37 NaF PET/CT scans with
1833 lesions (physician delineated and scored - 5-point scale)
nnUnet trained for lesion delineation and classification

Schott et al 2024, Phys Med Biol 69: 155015

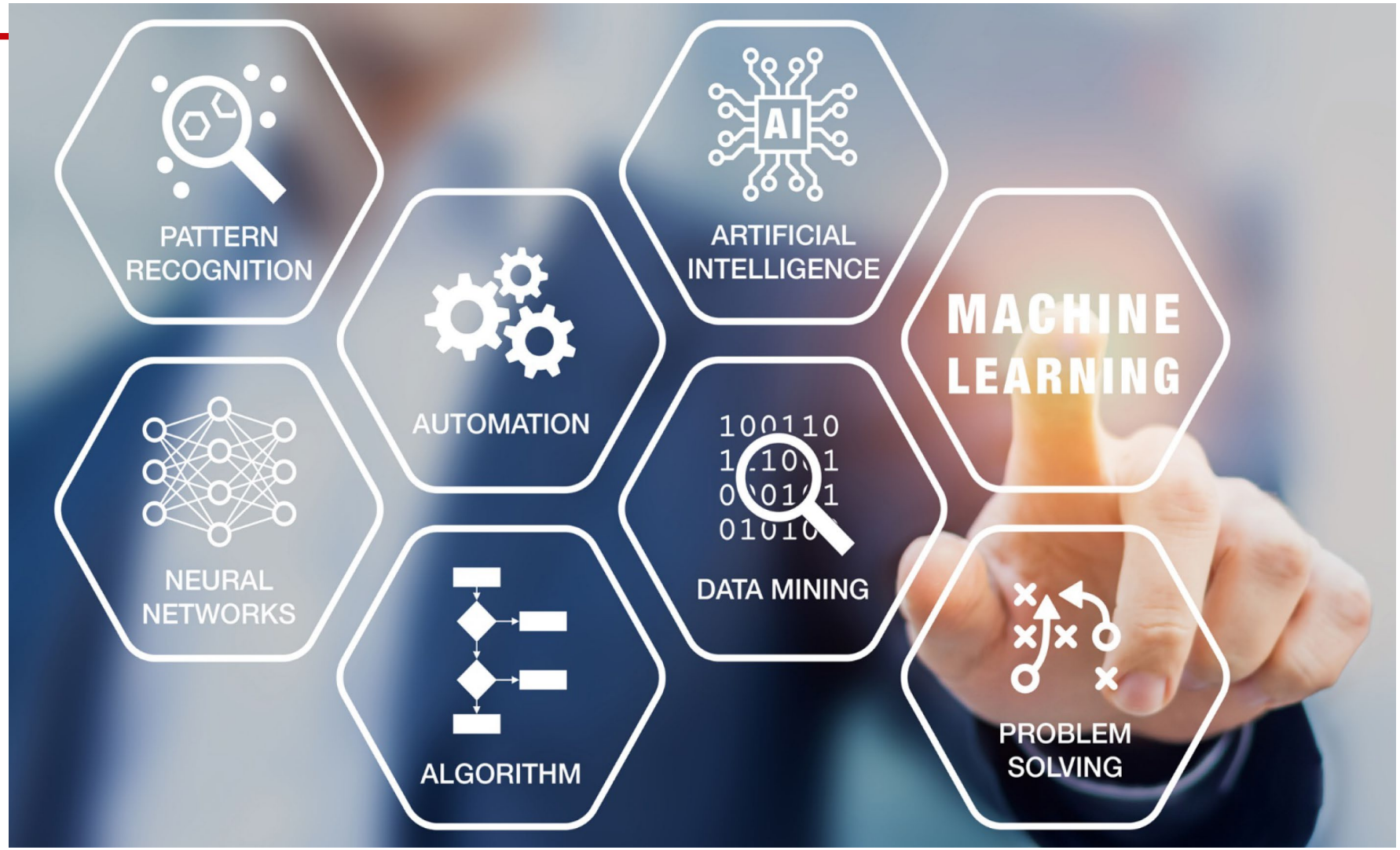
Summary: Safe deployment of AI



HOW CAN WE SAFELY DEPLOY AI-BASED QIB?

- Regulatory initiatives

FDA's concerns



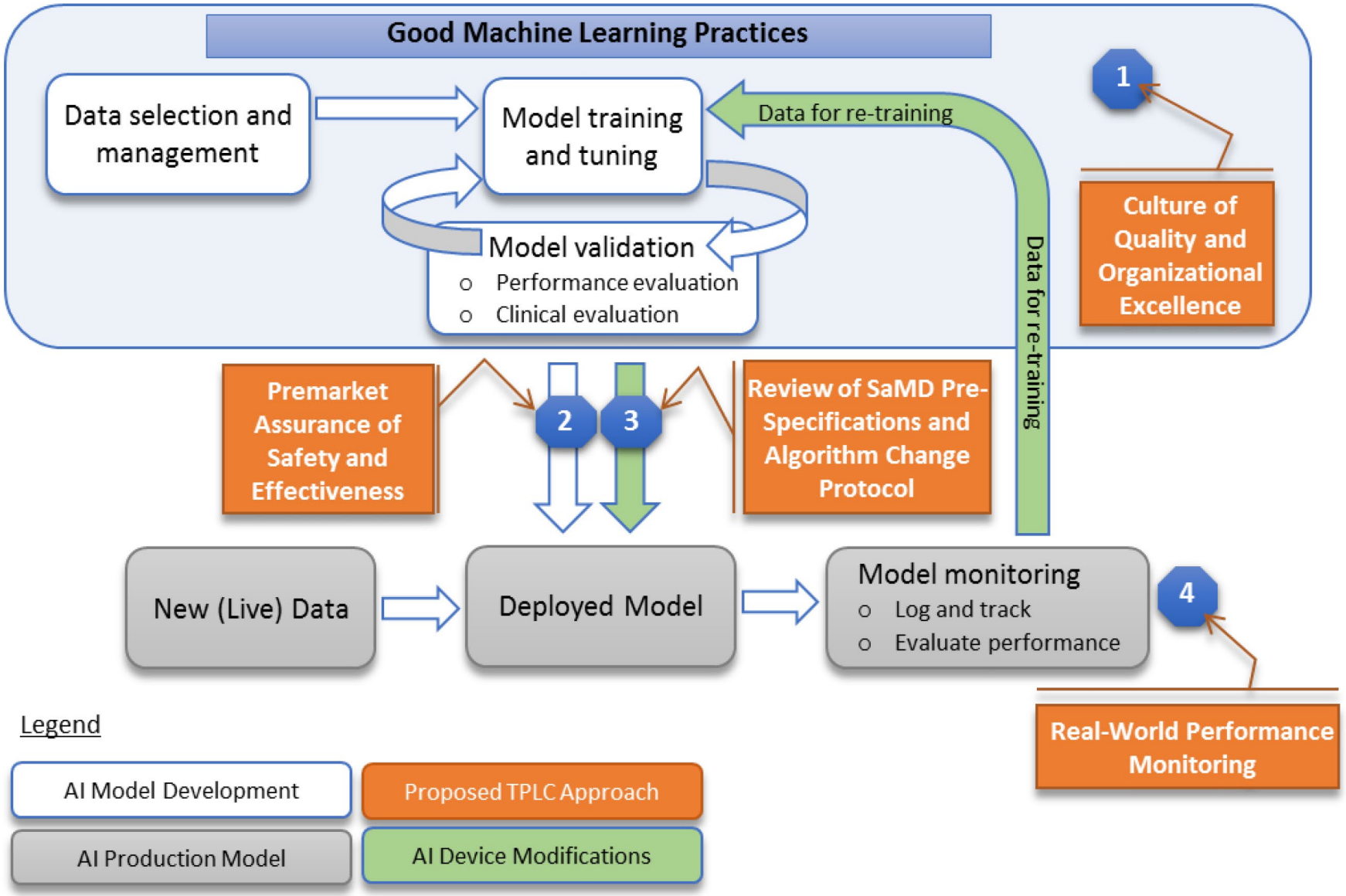
Proposed Regulatory Framework for Modifications to AI/ML-Based Software as a Medical Device (SaMD)

Concerns with AI/ML software

- To address the critical question of when a continuously learning AI/ML SaMD may require a premarket submission for an algorithm change, we were prompted to reimagine an approach to premarket review for AI/ML-driven software modifications.

- Such an approach would need to maintain reasonable assurance of safety and effectiveness of AI/ML-based SaMD, while allowing the software to continue to learn and evolve over time to improve patient care.

Good Machine Learning Practices (GMLP)

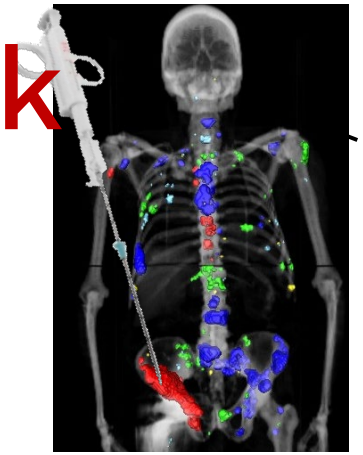


School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MAI

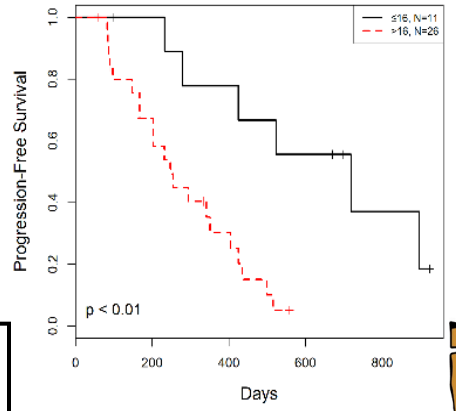
FDA Action plan

- Further developing the proposed regulatory framework, including *predetermined change control plan* (for software's learning over time)
- Supporting the development of good machine learning practices to evaluate and improve machine learning algorithms
- Fostering a patient-centered approach, including device transparency to users
- Developing methods to evaluate and improve machine learning algorithms
- Advancing real-world performance monitoring pilots.

Optimizing AI-QIBs for the task



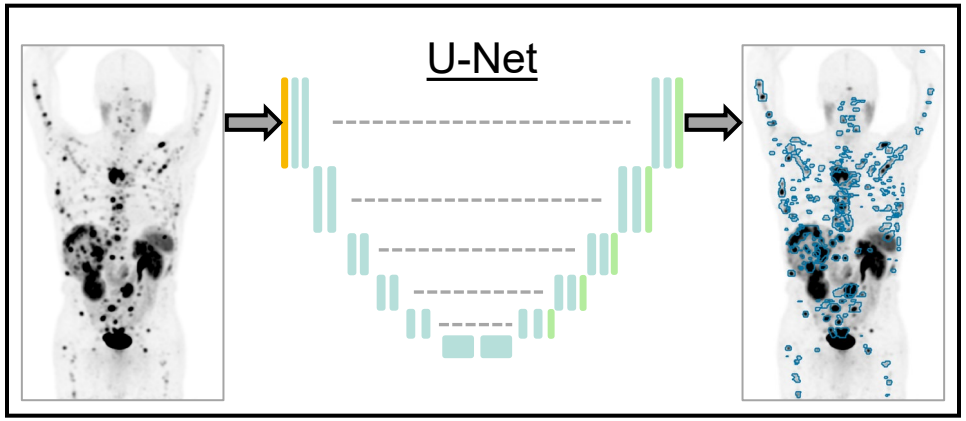
Treatment Interventions



Schott et al., 2023

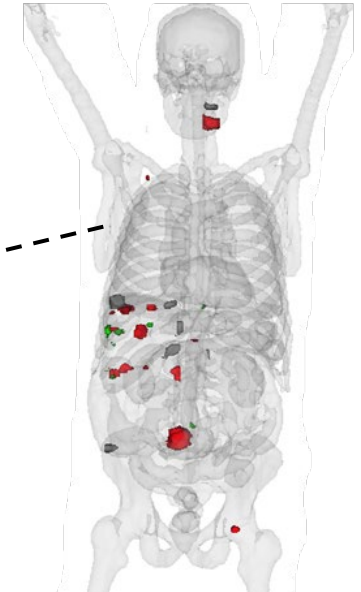
Predictive Models

Automatic Delineations



Follow-up Assessments

Response Map



Summary

- Quantitative Imaging Biomarkers (QIBs) are critical:
  - Predicting response during treatment (treatment response biomarkers)
  - Predicting patient outcome (surrogate endpoints)

- AI-based QIBs are essential:
  - Assessment of each individual lesion response (metastatic disease)
  - Modeling complex relationship to predict risks and benefits

- Safe deployment of AI-supported Precision Medicine is critical:
  - Out Of Domain (OOD) and In-Domain (ID) uncertainties
  - Optimizing AI for the task

Thanks to:

Research groups:

University of Wisconsin, WI, USA

University of Ljubljana, Slovenia

Collaborators: University of Wisconsin (USA), University of Ljubljana (SLO), OIL, UKCL, AIQ Solutions

Funding: NIH (R01, P30, P50, SBIR), ARIS

School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MADISON

Thank you
for your attention

School of Medicine
and Public Health
UNIVERSITY OF WISCONSIN–MADISON