WashU Medicine

Quality assurance and monitoring of clinical AI systems

Geoff Hugo, Ph.D.

gdhugo@wustl.edu

© 2026 WashU Medicine

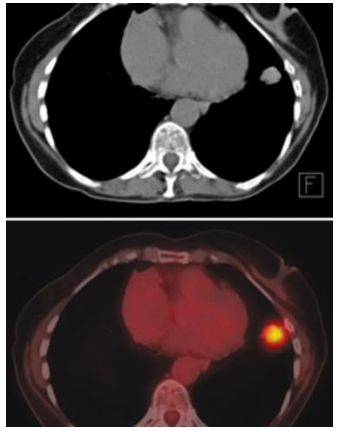
Disclosures

- **Employee**: Washington University
- **Research Grants**: NIH, Siemens, Varian Medical Systems
- **Speaking / consulting**: Varian, unrelated to present work.
- **Licensing**: Varian, unrelated to present work



Symptoms / Screening

Wk to Mo



Diagnosis / Staging

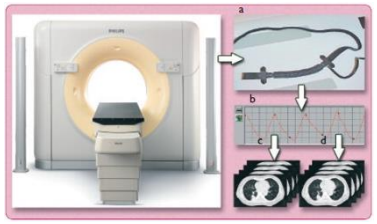
Wks

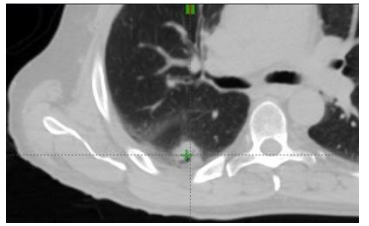


Consultation /
Treatment Selection /
Trial participation

~4d

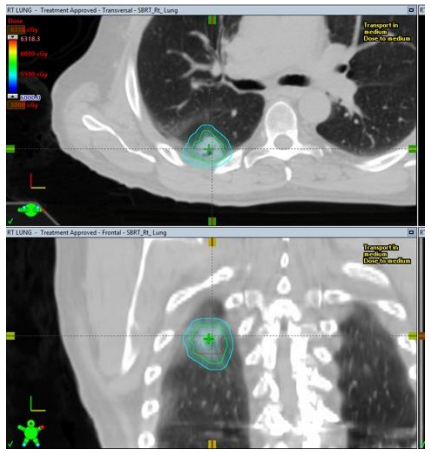






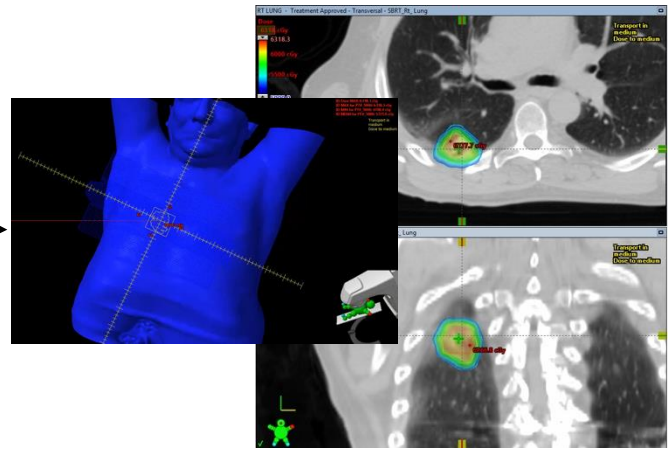
Imaging / Simulation

~3d



Segmentation
(Contouring)

~2d



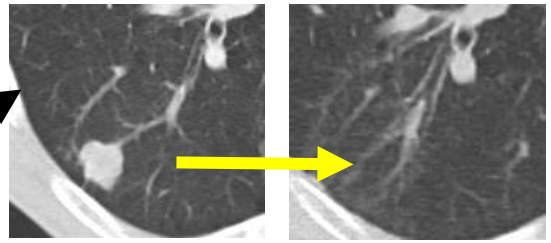
Treatment Planning

~1d



Delivery

~Wks to yrs



Cancer Outcomes

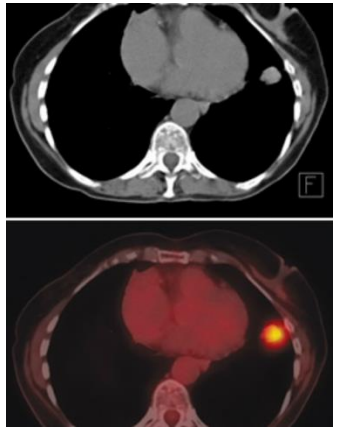


Toxicity



Earlier detection (Screening Adherence, Internet Searches)

Wk to Mo



AI reading, radiomics

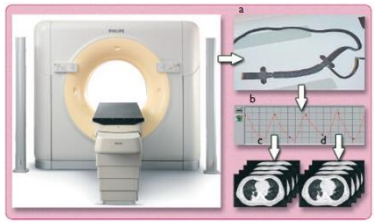
Wks

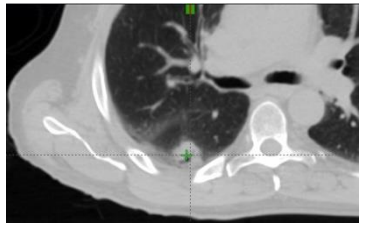


Guideline adherence vs. “personalized” treatment

~4d

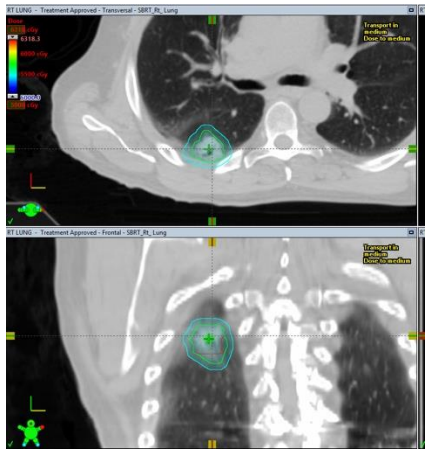






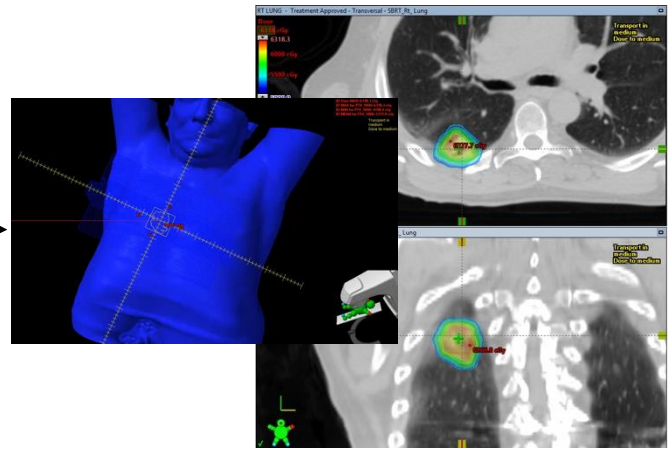
Optimized imaging, CDS for simulation selection

~3d



Auto-contouring, auto-contouring QA

~2d



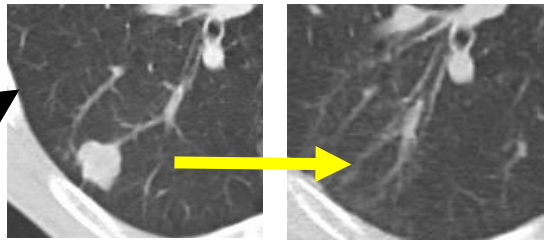
Auto-planning, auto-planning QA

~1d



CDS for delivery device (i.e. IMRT vs. protons)
Adaptive RT

~Wks to yrs



Auto-data capture / analysis

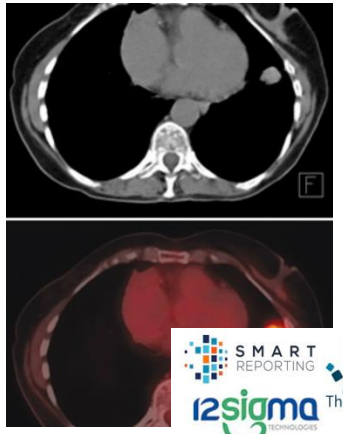


Auto-data capture



Earlier detection (Screening Adherence, Internet Searches)

Wk to Mo



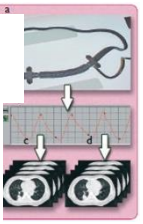
AI reading

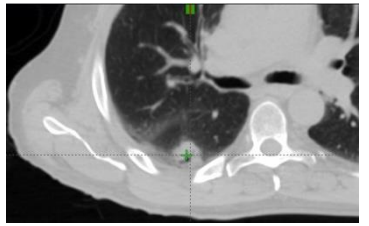
Wks



~4d



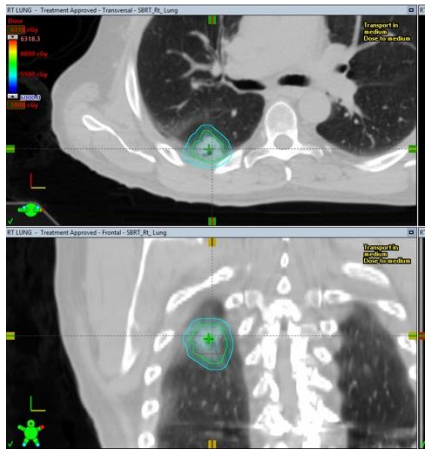




zed imaging, CDS for
nulation selection

via Curt Langlotz, @curtlanglotz

SMART REPORTING · Medexprim · 商汤 sensetime · median · OSIMIS · QLARITY IMAGING · image biopsy lab · 12sigma · Therapixel · SEMA · CuraCloud · enlitic · QMENTA · PaxeraHealth · qure.ai · quibim · zebra · ai analysis inc. · Blackford · CureMetrix · JLK INSPECTION · WEKA.IO · teraricon · VAREX IMAGING · ADVANTIS · ANYMEDI · BOLD BRAIN VENTURES · DNP · FLYWHEEL · INTELLIGENT ULTRASOUND · PHILIPS Healthcare · CORTECHS Labs · aidence · CAIDE Systems · DDH · VARIAN · TOMTEC · imbio · PVmed · Sorna · aidoc · Cercare Medical · DEEP VOXEL · Yitu Healthcare · INCEPTO · Perspectum Diagnostics · SUBTLE MEDICAL · AIRS · circle · viz.ai · inferVISION · Quantib · DIA · COGNEX · DEEP NOID · FOVIA · Google Cloud Platform · KHEIRON · Rad AI · ONEMEDNET · IMAGING ANALYSIS · VUNO · ARTERYS · GE Healthcare · koios · OXIPIT · PURESTORAGE · BAYER · iCAD · COMBINOSTICS · DeepRadiology · HEALTHMYNE · MEDICVision IMAGING SOLUTIONS · RADLogics · VIDA · DeepTek · BioMind · onyx · SOPHiA · BALZANO · contextflow · densitas · IBM Watson · mirada · RADIOBOTICS · 数坤科技 SHUKUN · COLLECTIVE MINDS RADIOLOGY · 深睿医疗 deepwise · Fraunhofer MEVIS · icometrix · macadamian · CORELINE SOFT · HEARTVISTA · Lunit · mim · NUANCE

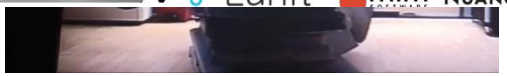


Auto-contouring, auto-contouring QA

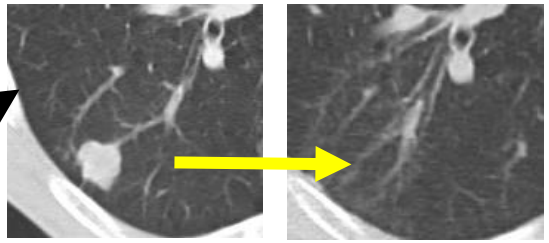
~2d



Auto-planning, auto-planning QA



CDS for delivery device (i.e. IMRT vs. protons) Adaptive RT



Auto-data capture / analysis

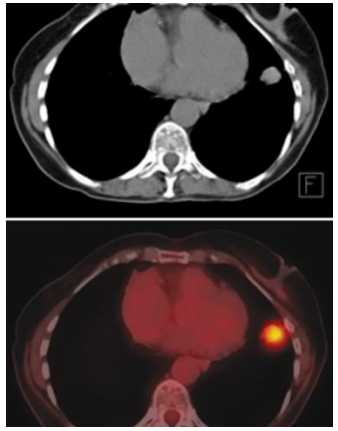


Auto-data capture



Earlier detection (Screening Adherence, Internet Searches)

Wk to Mo



AI reading, radiomics

Wks

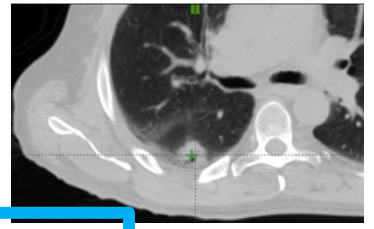


Guideline adherence vs. “personalized” treatment

~4d

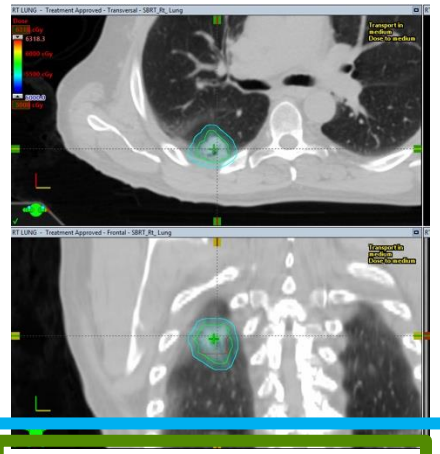






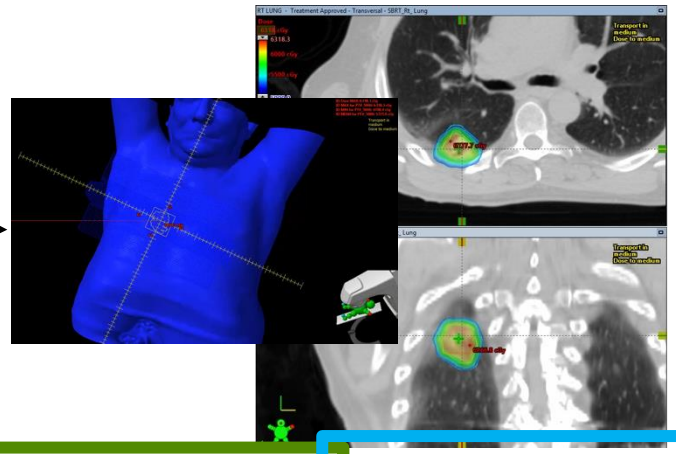
Optimized imaging, CDS for simulation selection

~3d



Auto-contouring, auto-contouring QA

~2d



Auto-planning

auto-planning QA

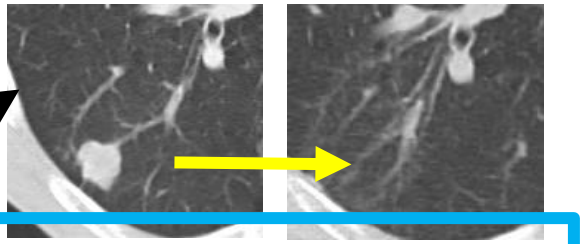
~1d



CDS for delivery device (i.e. IMRT vs. [illegible])

Adaptive RT

~Wks to yrs



Auto-data capture / analysis



Auto-data capture

What is Medical Physics?

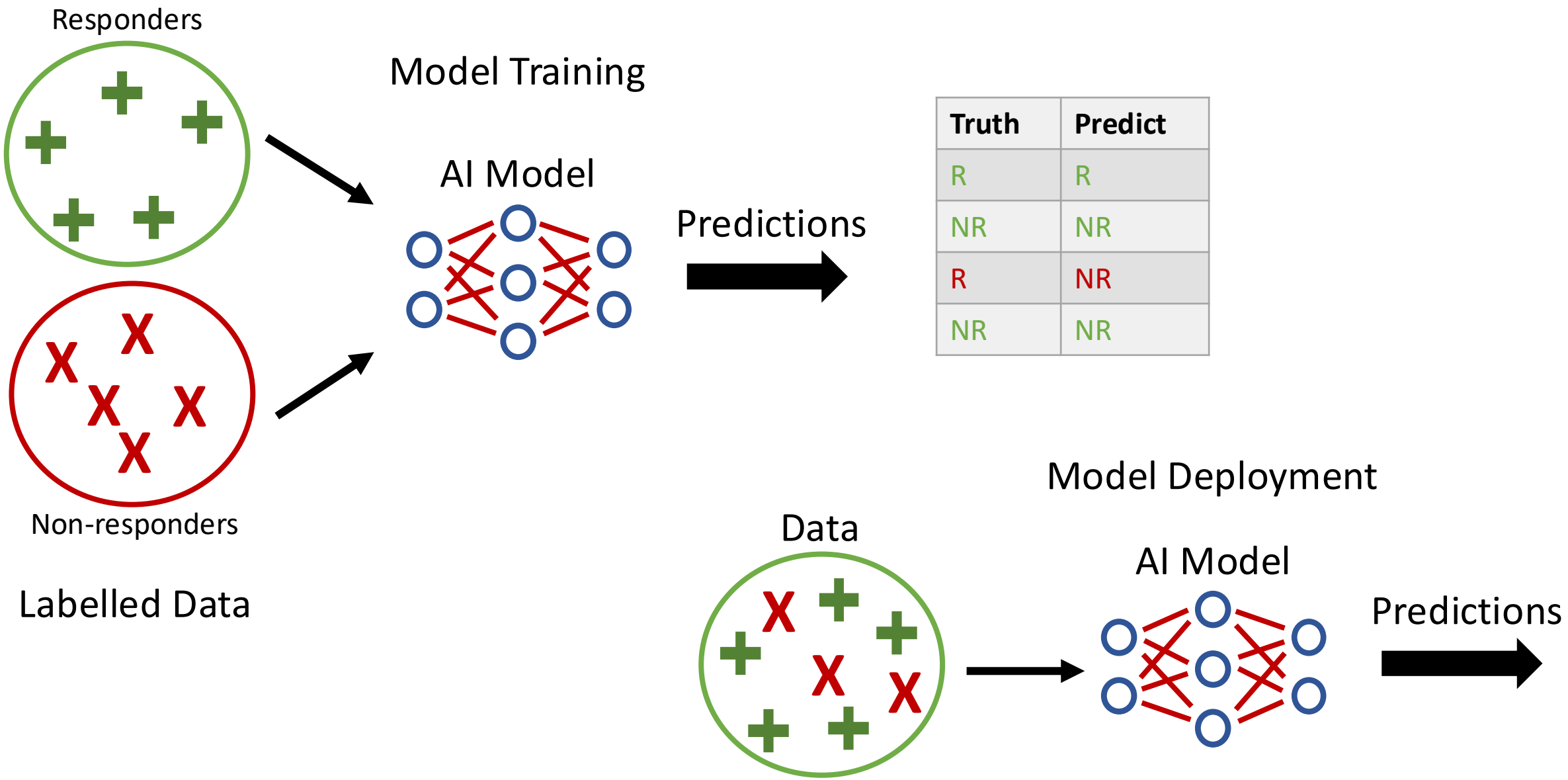


Dr Sonja Dieterich
@SonjaDieterich

Following

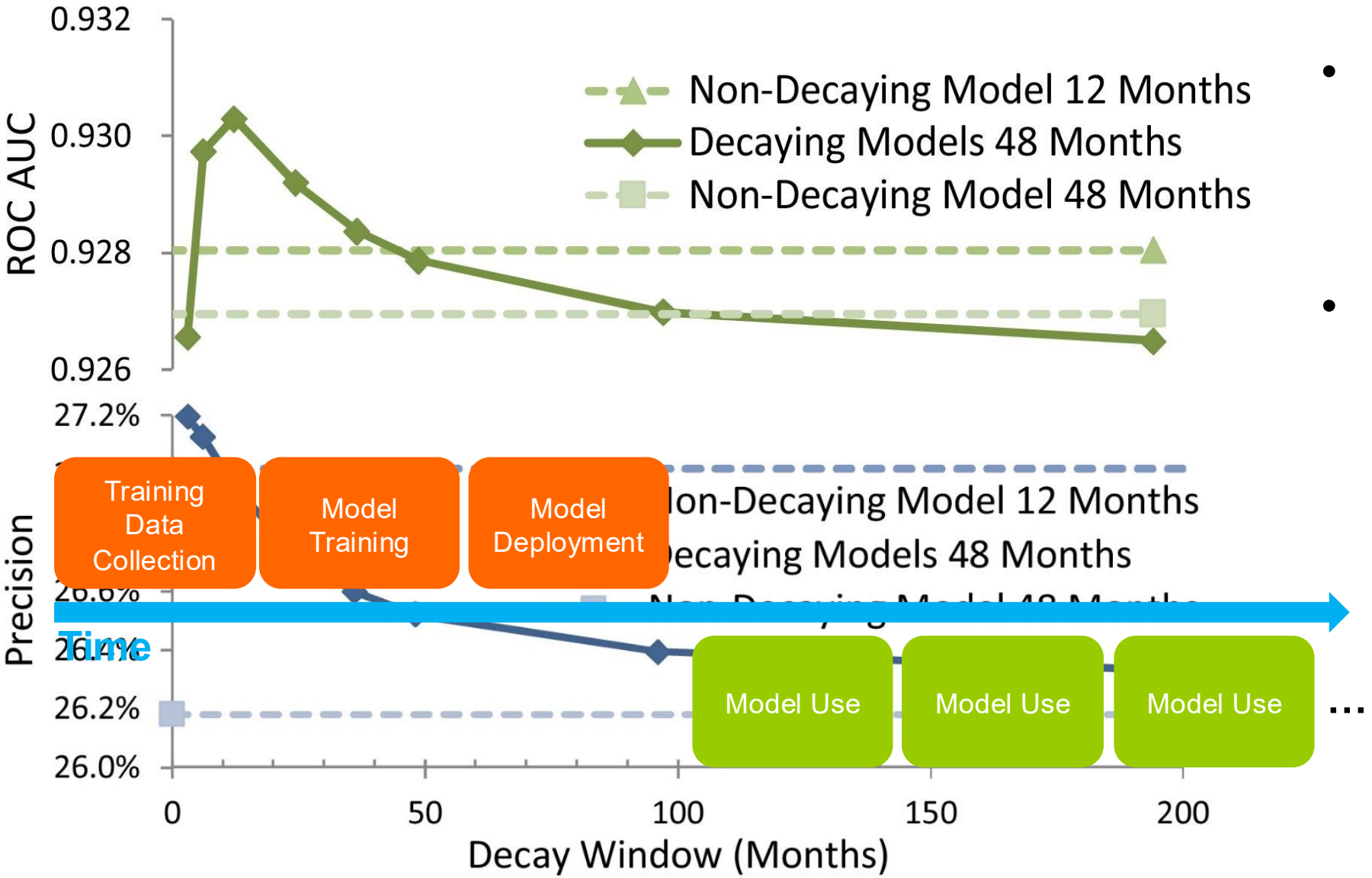
I am a pharmacist for radiation.
#MyScienceIn6words

How does AI learn?



Truth	Predict
R	R
NR	NR
R	NR
NR	NR

The Problem: Supervised Learning



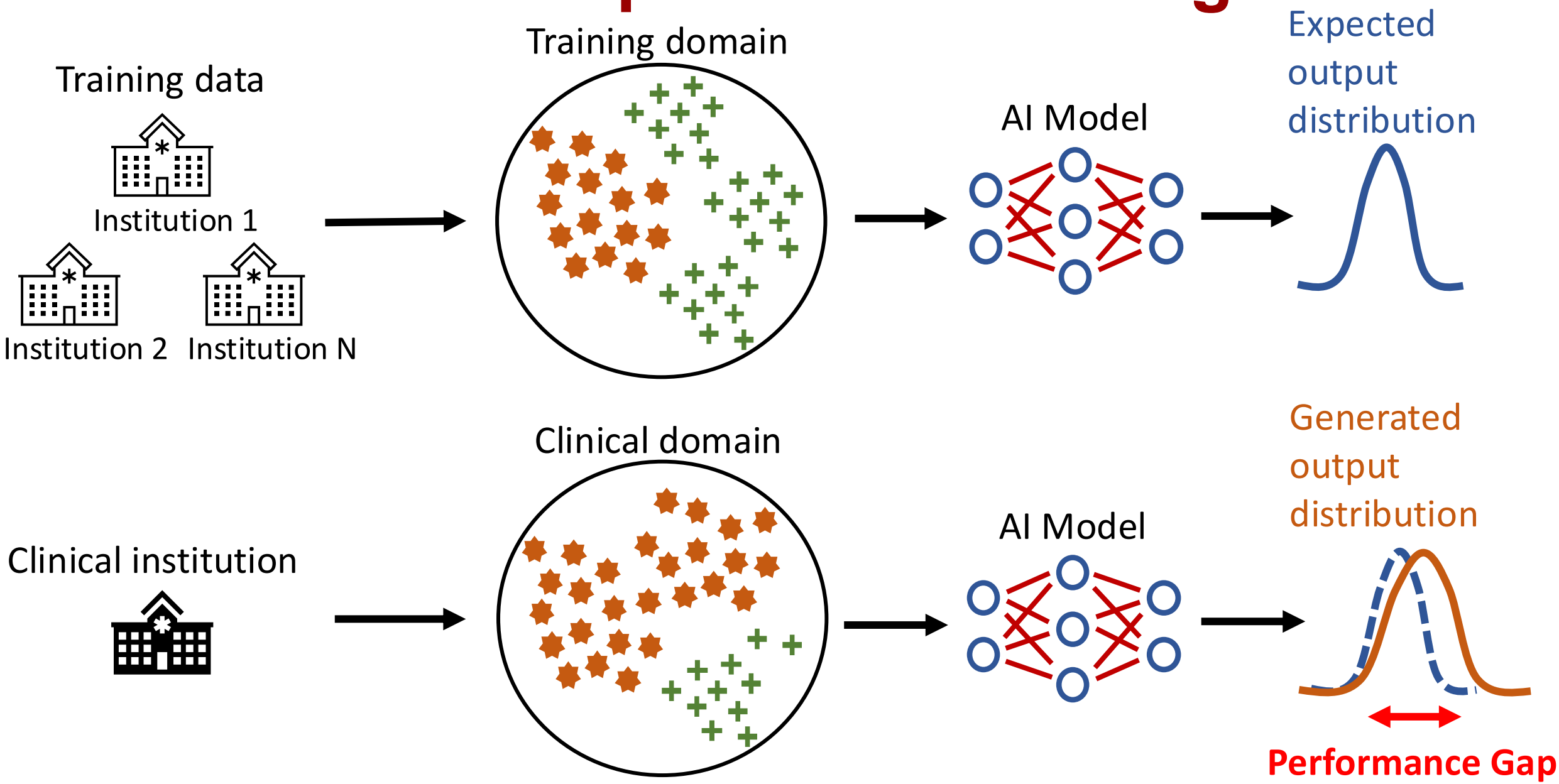
- AI model performance decays with a ”half-life” of 6 months
- Fundamental problem due to how these models are trained and then deployed

ELSEVIER
International Journal of Medical Informatics
Volume 102, June 2017, Pages 71-79

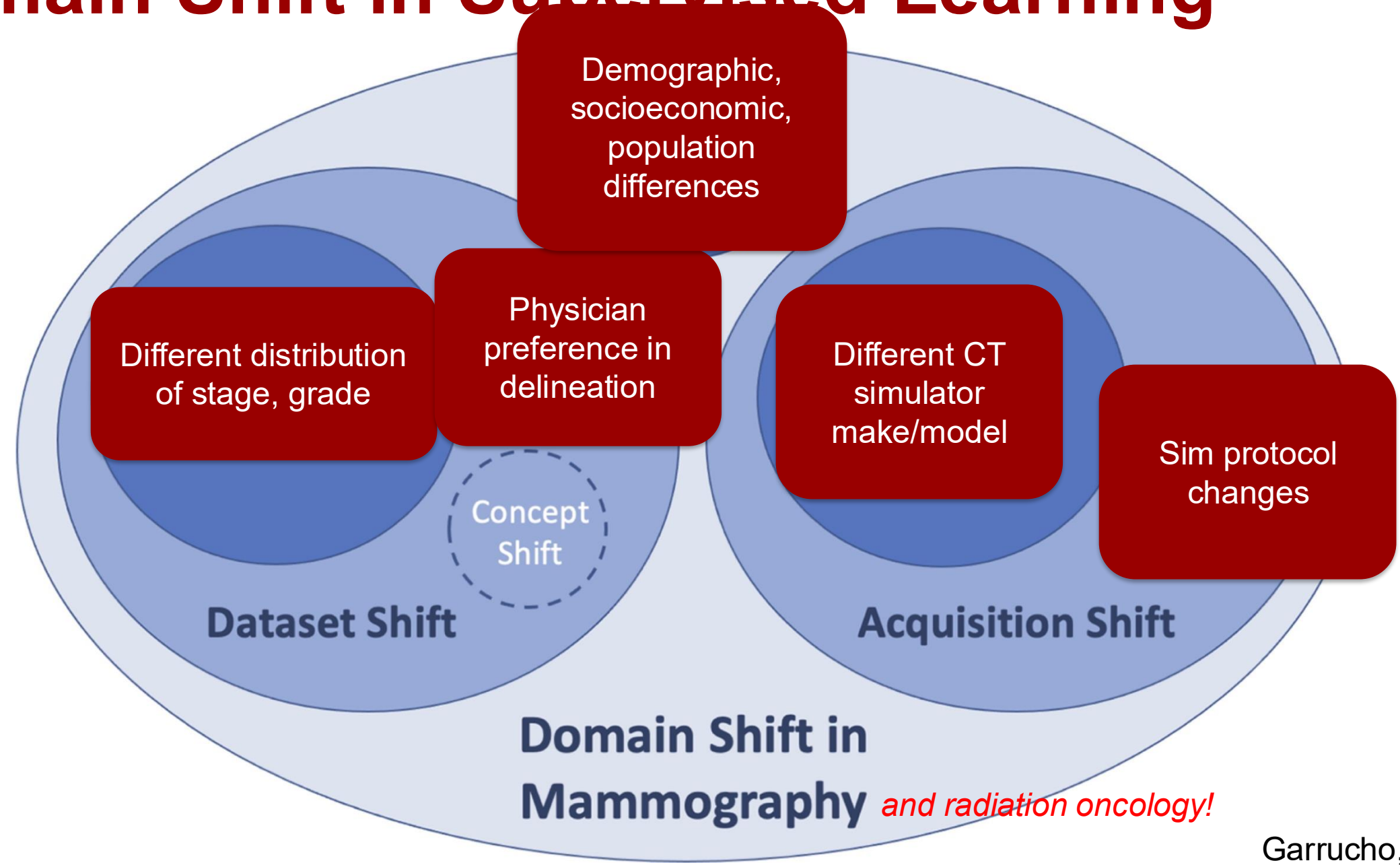
Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets

Jonathan H. Chen [a], Muthuraman Alagappan [d], Mary K. Goldstein [b c], Steven M. Asch [a f], Russ B. Altman [a e]

Domain Shift in Supervised Learning



Domain Shift in Supervised Learning

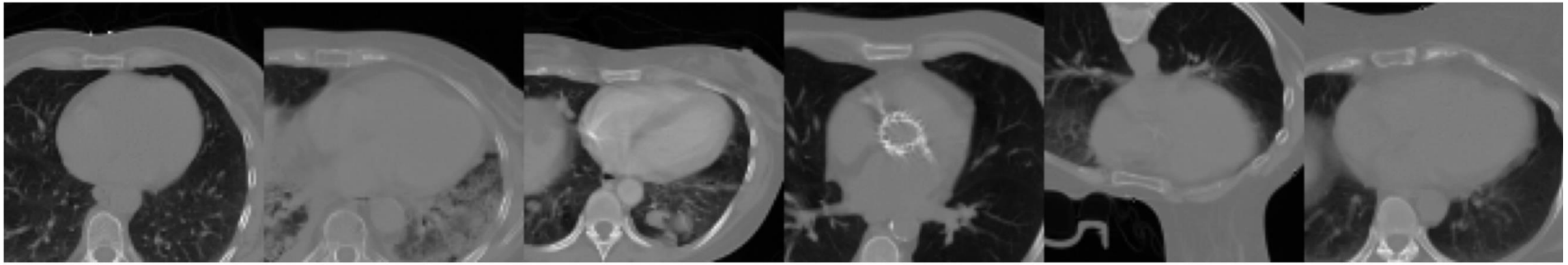


Garrucho, AIM 2022

Study Goals

- Investigate sensitivity of AI autosegmentation to domain shift
- Understand types of domain shift that affect autosegmentation performance
- Build approaches to detect domain shift
- Eventually: correct for domain shift?

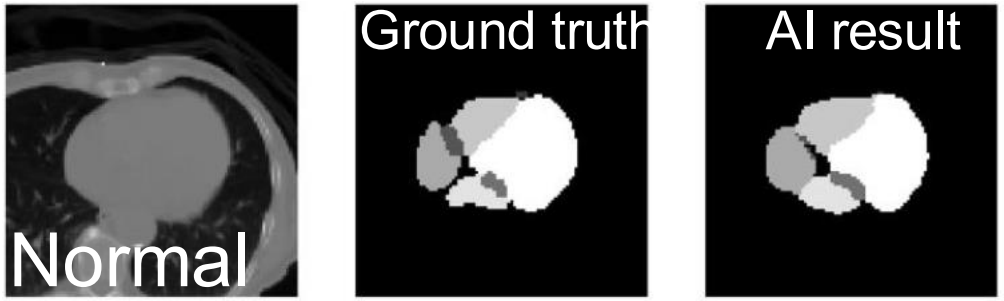
Benchmark Dataset



- 241 breast cancer RT patients
- 60 held out for test dataset
- Heart substructures delineated

Test AI autosegmentation models

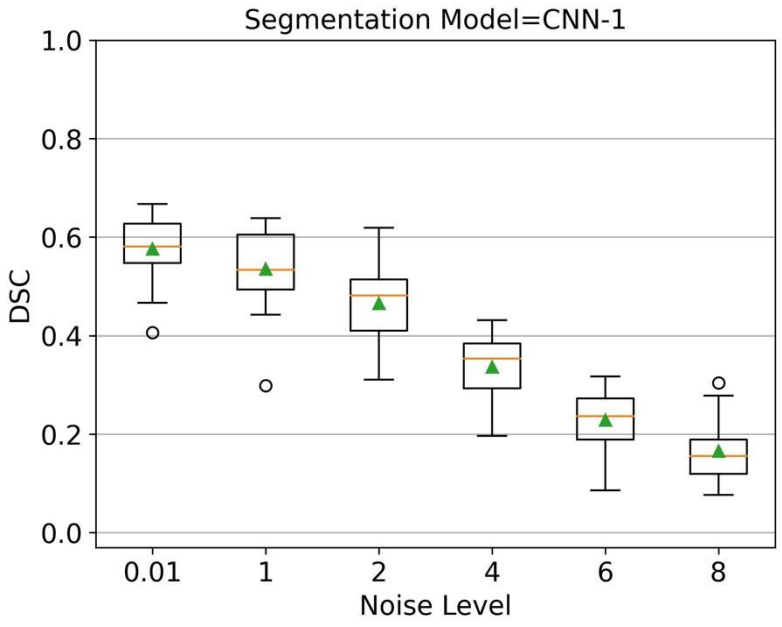
- 18 various convolutional neural network-based (CNN) autosegmentation models
  - U-net based
  - Developed and tested on a separate, in-house dataset
  - Various loss functions, network depth, data augmentation / training strategies
- 1 ViT (vision transformer) trained on same data as CNNs
- 1 Commercial heart substructure AI model (Siemens Syngo.via) – designed for contrast CT

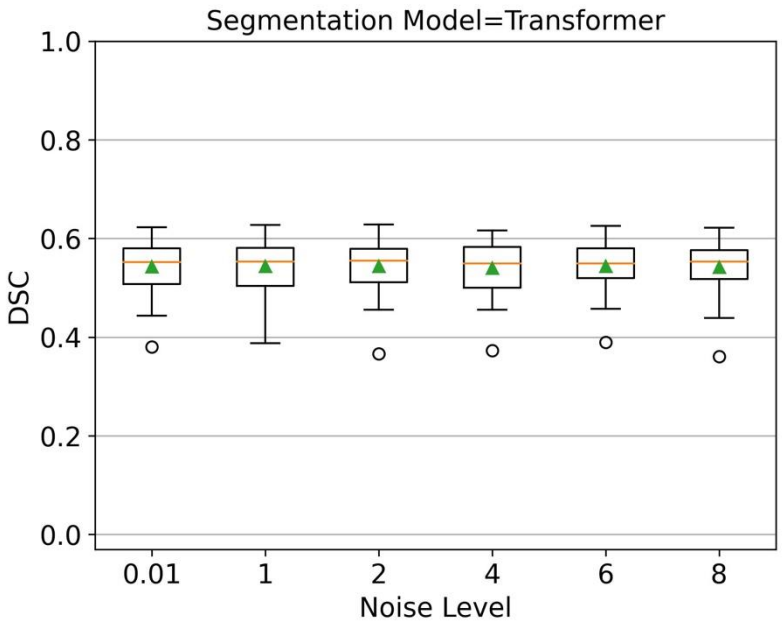


Results – Scanner Protocol Variation

- **Image noise** reduced model segmentation accuracy significantly for CNN models (p=0.01), but not for the Transformer
- No significant effect (p=0.42) from **slice thickness, FOV, and reconstruction kernel** on segmentation model performance. (not shown)

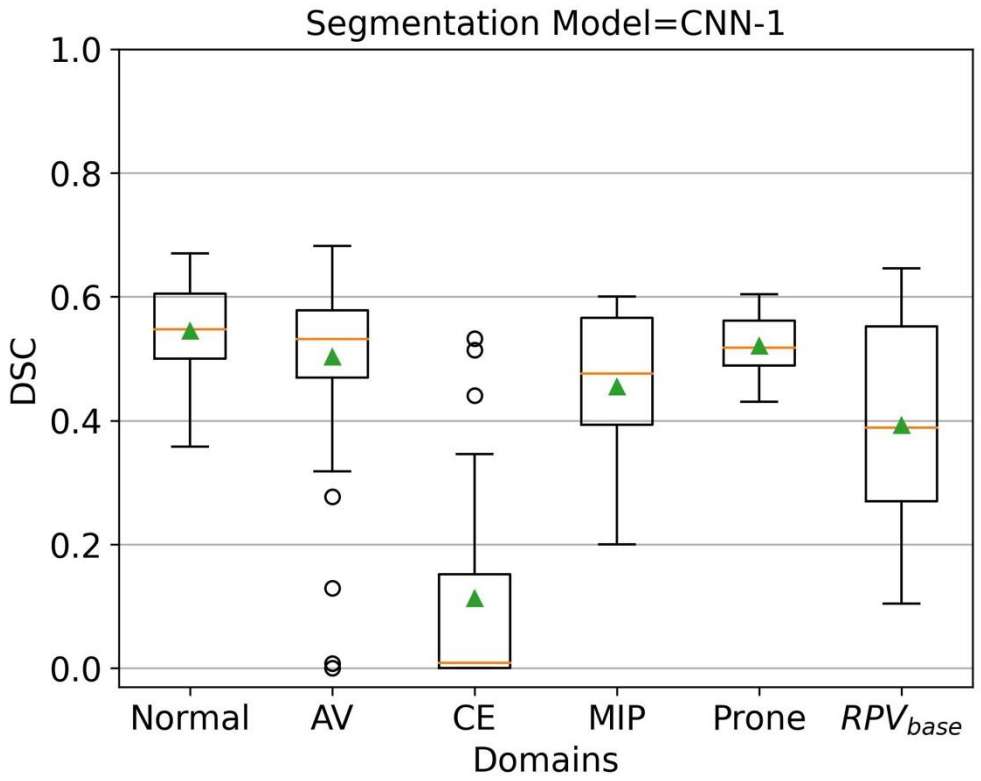
Image Noise

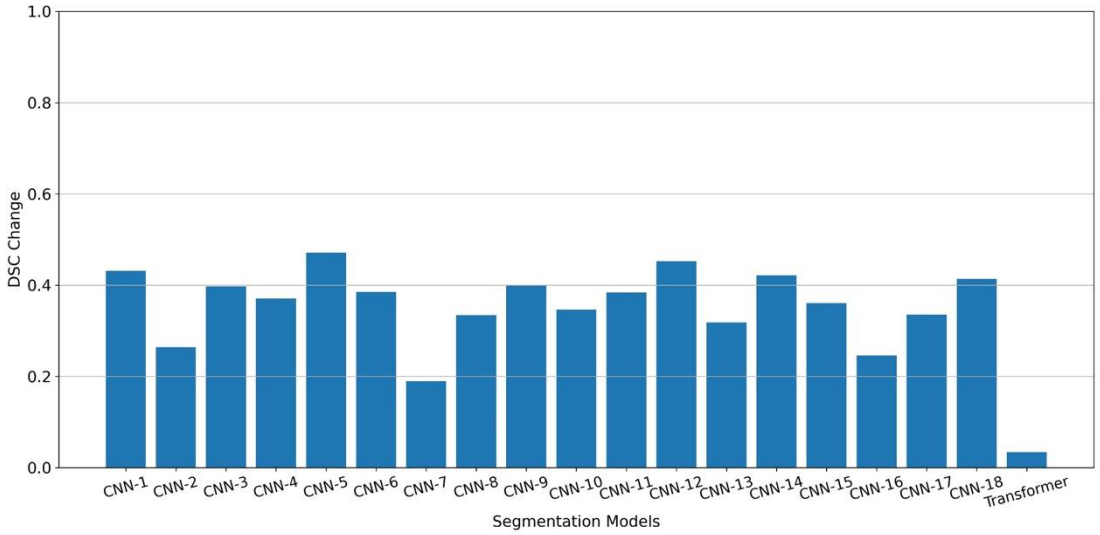




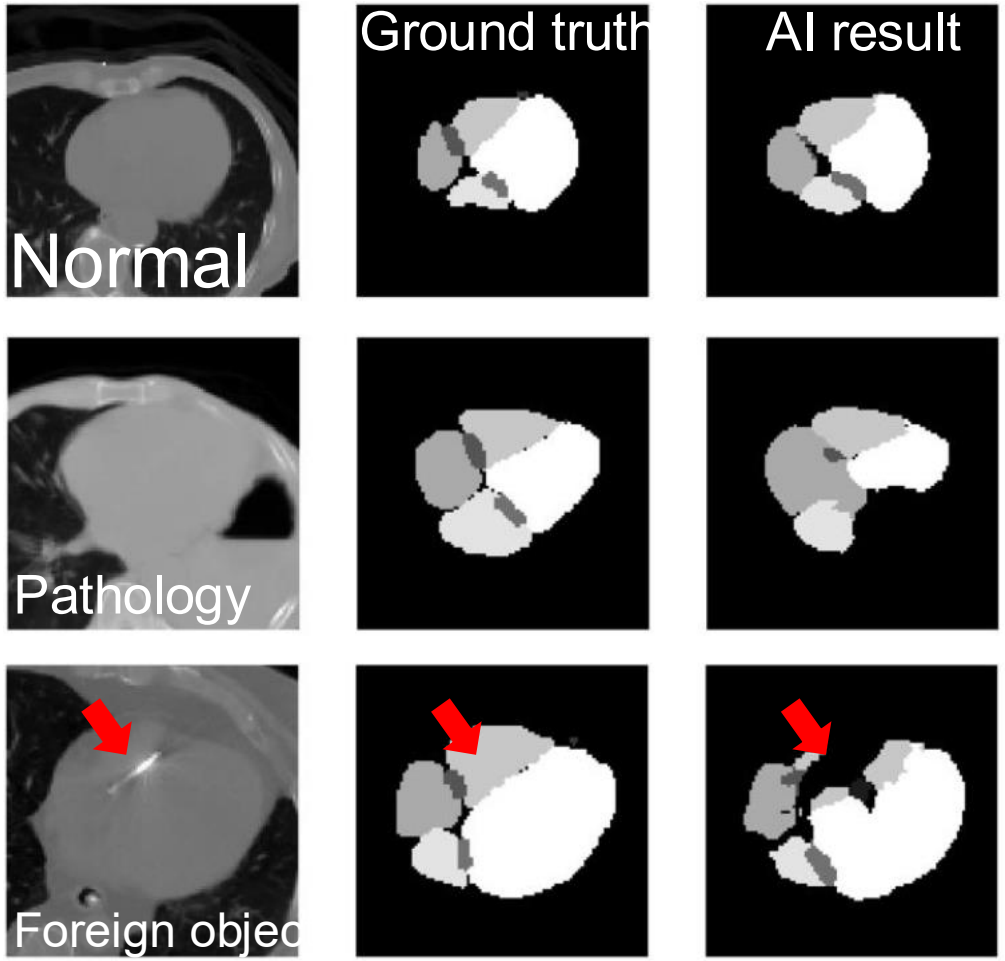
Results – Other Domain

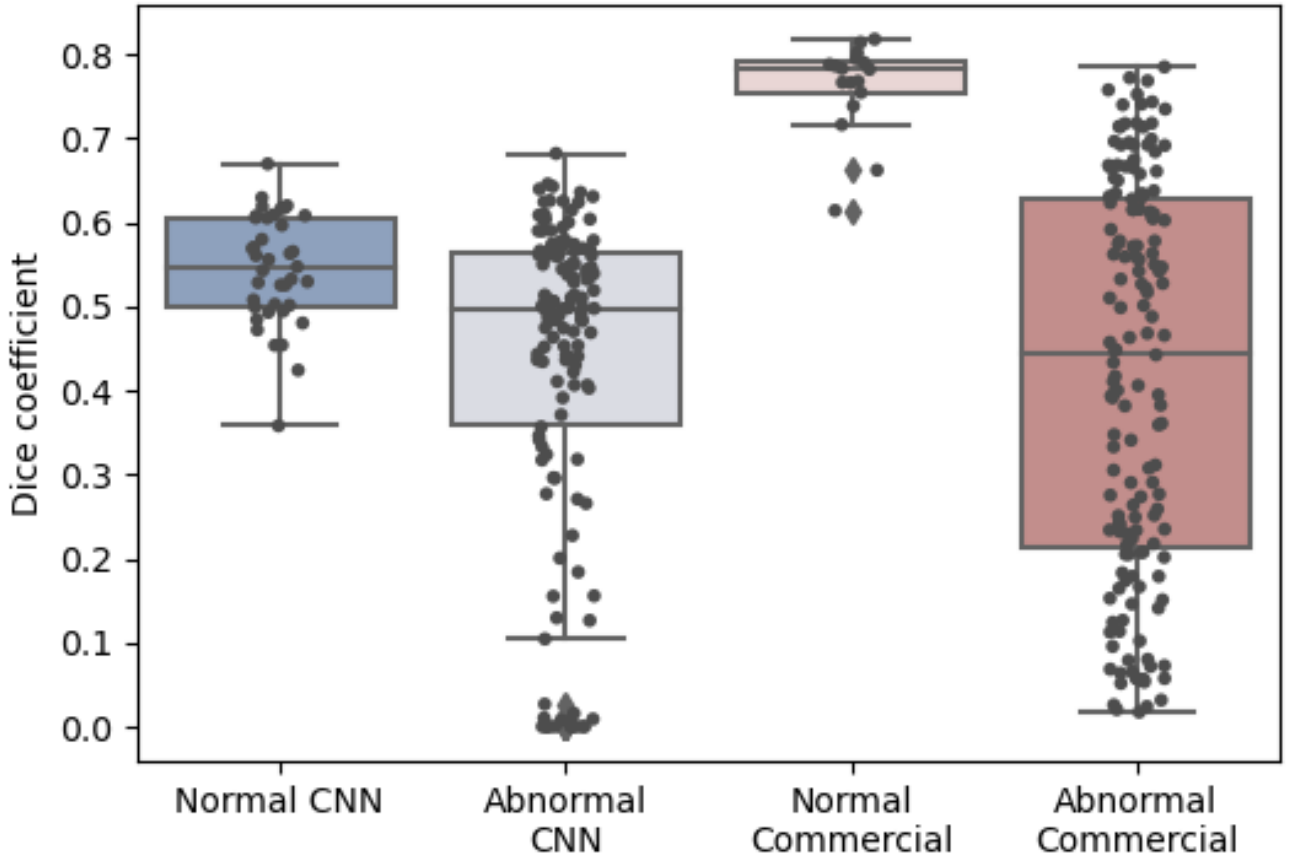
- CNN models showed significant drop ($p<1e-3$) on the anatomical variation and contrast domains along with image noise (RPV_{base})
- Transformer performance was more stable





Results – Commercial vs. In House





Domain Shift in the Clinic

PAPER • OPEN ACCESS

Performance deterioration of deep learning models after clinical deployment: a case study with auto-segmentation for definitive prostate cancer radiotherapy

Biling Wang[1,3,5], Michael Dohopolski[1,2,5], Ti Bai[1,2], Junjie Wu[1,2], Raquibul Hannan[1,2], Neil Desai[1,2], Aurelie Garant[1,2], Daniel Yang[1,2], Dan Nguyen[1,2], Mu-Han Lin[1,2] Show full author list

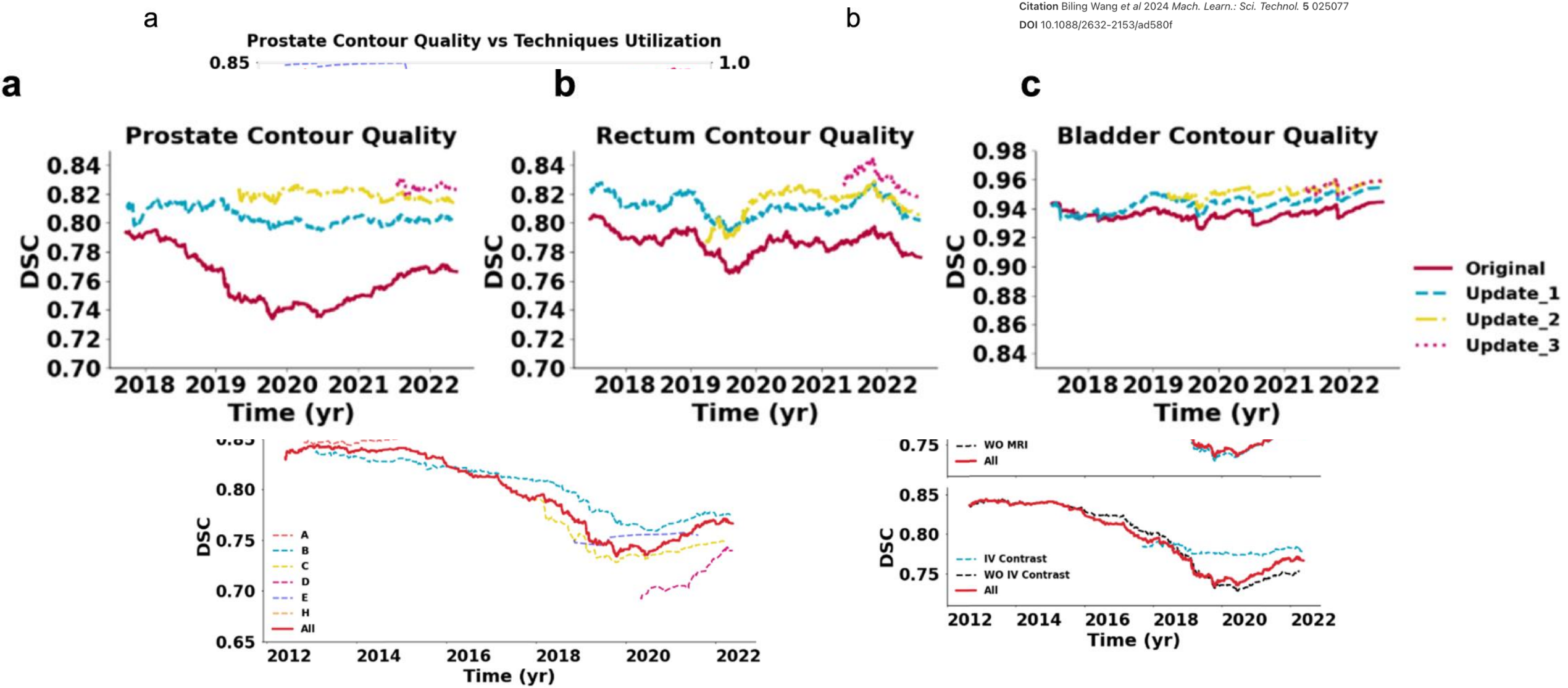
Published 25 June 2024 • © 2024 The Author(s). Published by IOP Publishing Ltd

Machine Learning: Science and Technology, Volume 5, Number 2

Focus on Machine Learning Models in Medical Imaging

Citation Biling Wang *et al* 2024 *Mach. Learn.: Sci. Technol.* **5** 025077

DOI 10.1088/2632-2153/ad580f

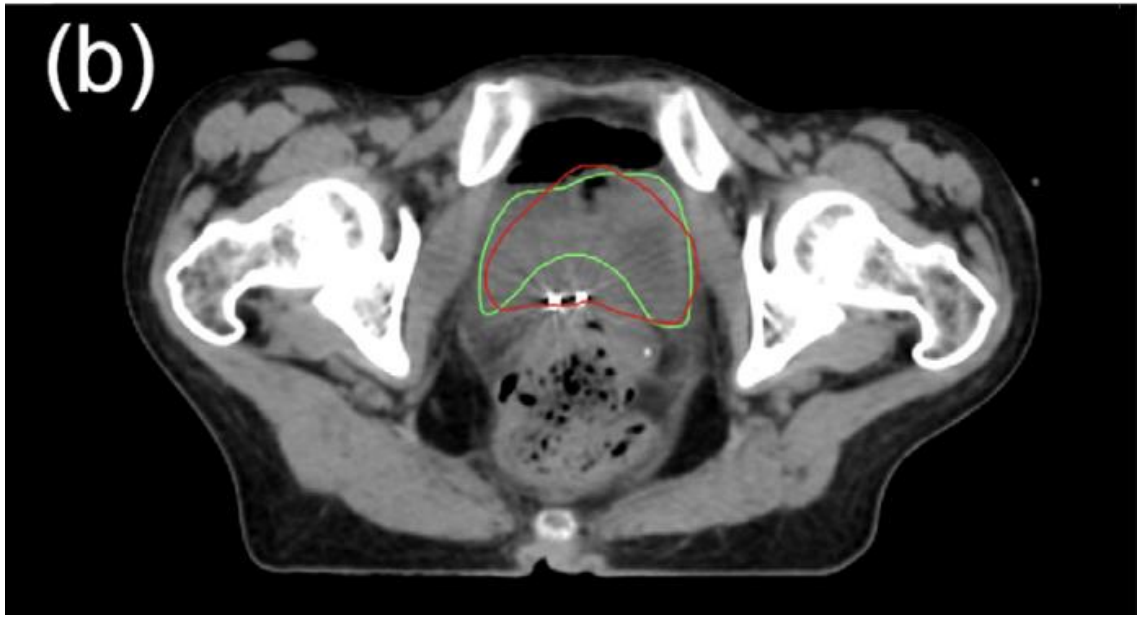


What if we don’t have access to change the model?

Can we monitor and flag domain shift?

Monitoring and Detection – Redundancy

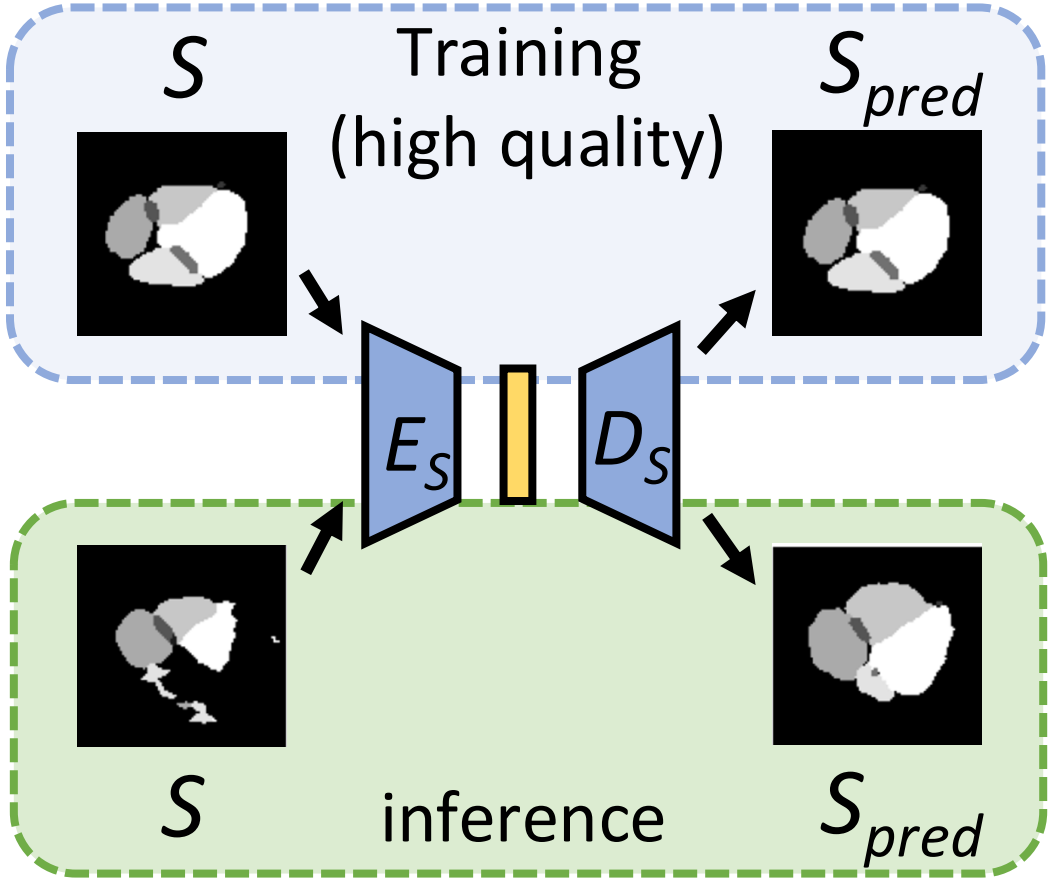
- Use a reference and secondary AI autosegmentation system
- SVM used to build a model to predict acceptable / unacceptable contours by contour overlap metrics
- Accuracy of 0.9 for cervical cancer contours



Rhee et al. JACMP 2022 https://doi.org/10.1002/acm2.13647

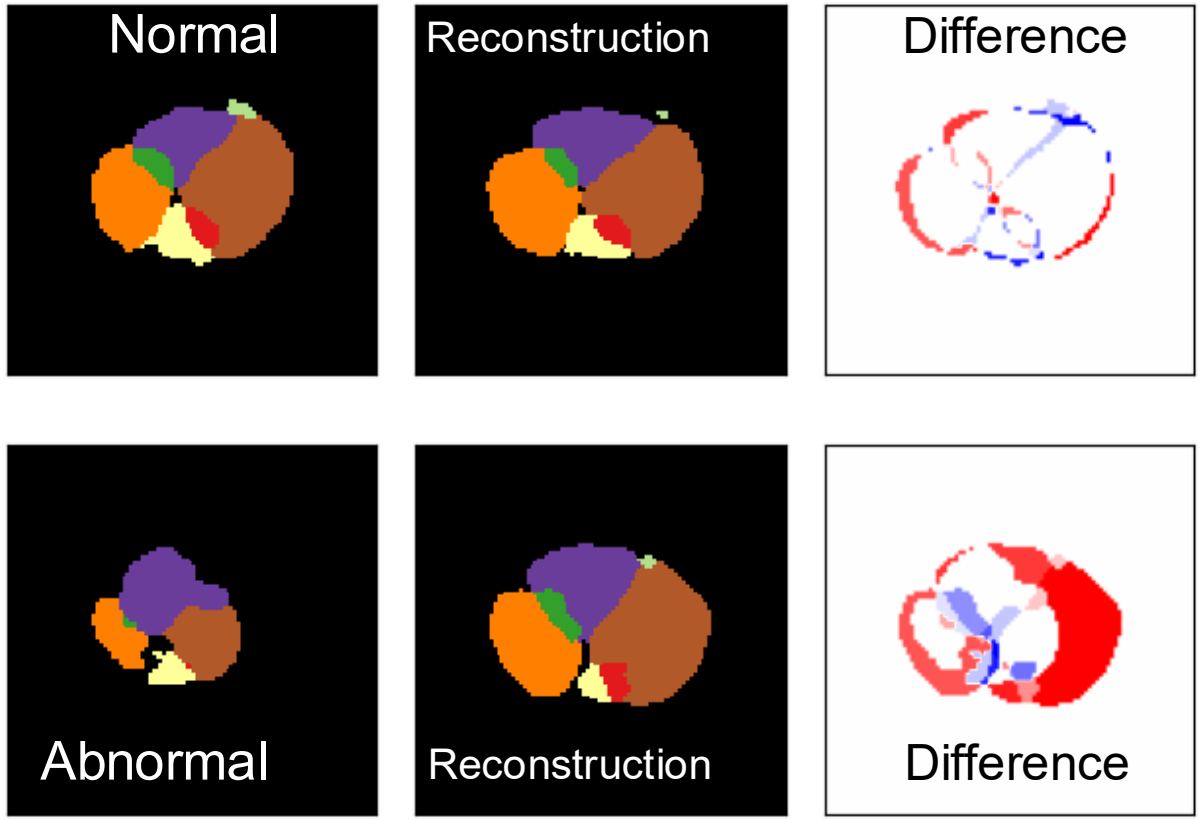
Shape Quality Estimator

- "Turn the tables" on sensitivity of AI to domain shift
- Train a small model (autoencoder) on clinically-acceptable data / results of an AI system
- Autoencoder is self-supervised: tries to compress and reconstruct input itself
- Out of domain data is poorly reconstructed

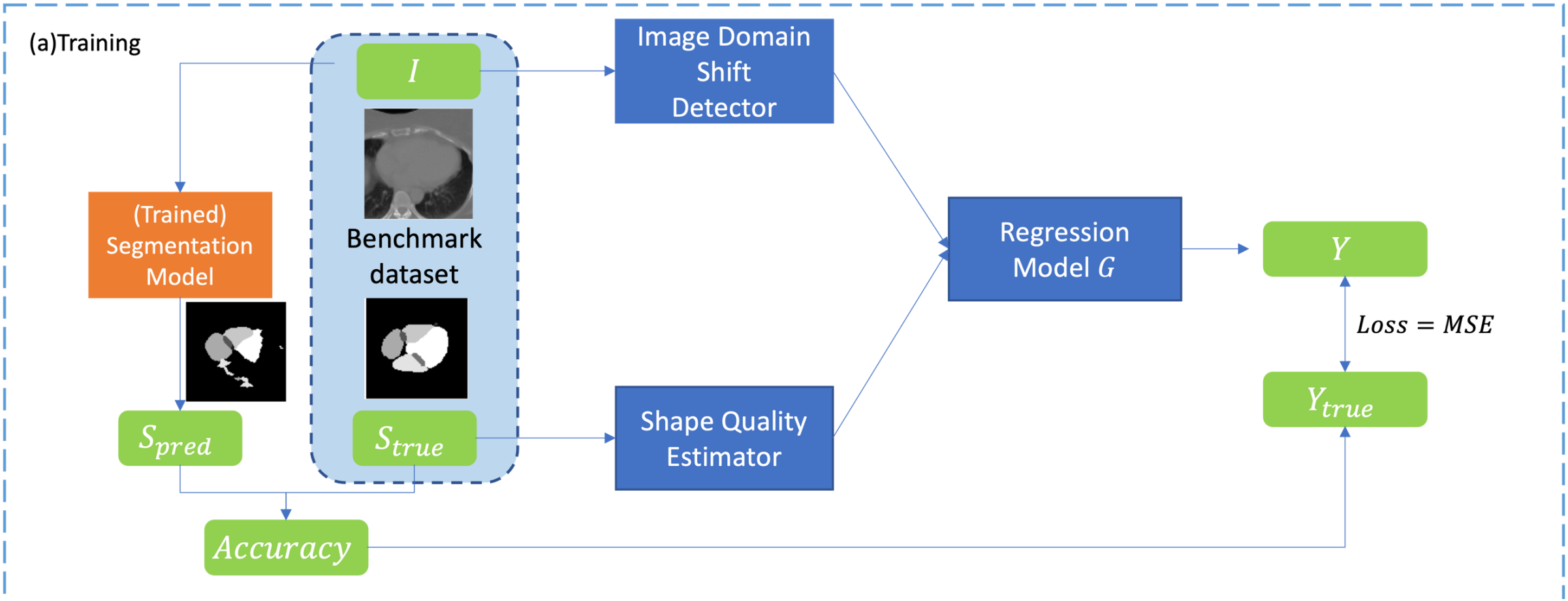


Shape Quality Estimator

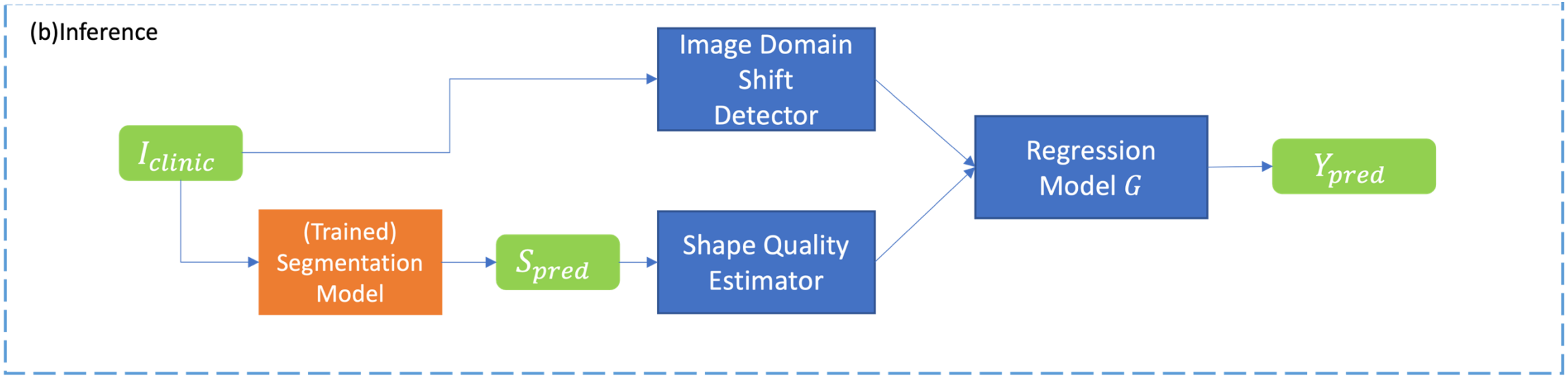
- “Turn the tables” on sensitivity of AI to domain shift
- Train a small model (autoencoder) on clinically-acceptable outputs of an autosegmentation system
- Autoencoder is self-supervised: tries to compress and reconstruct input itself
- Out of domain data is poorly reconstructed



Monitoring and Detection

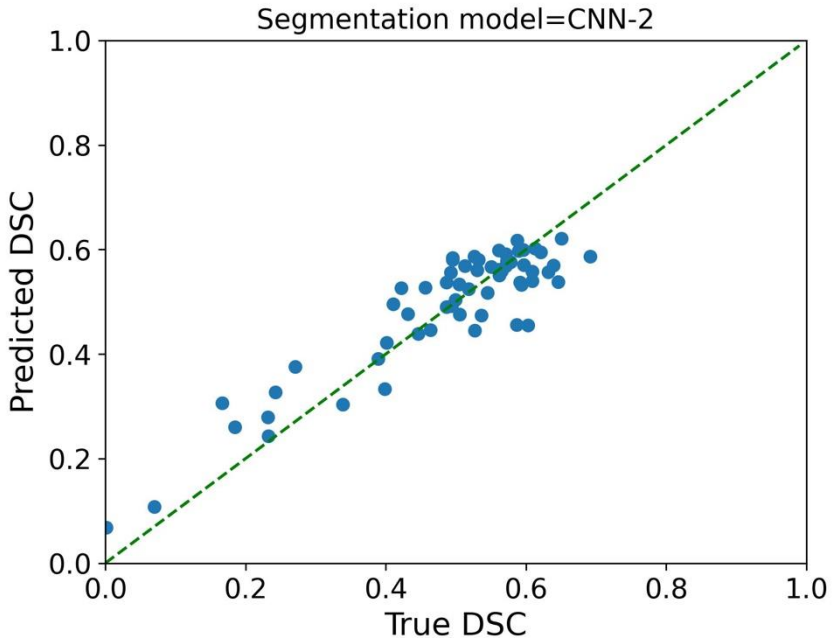


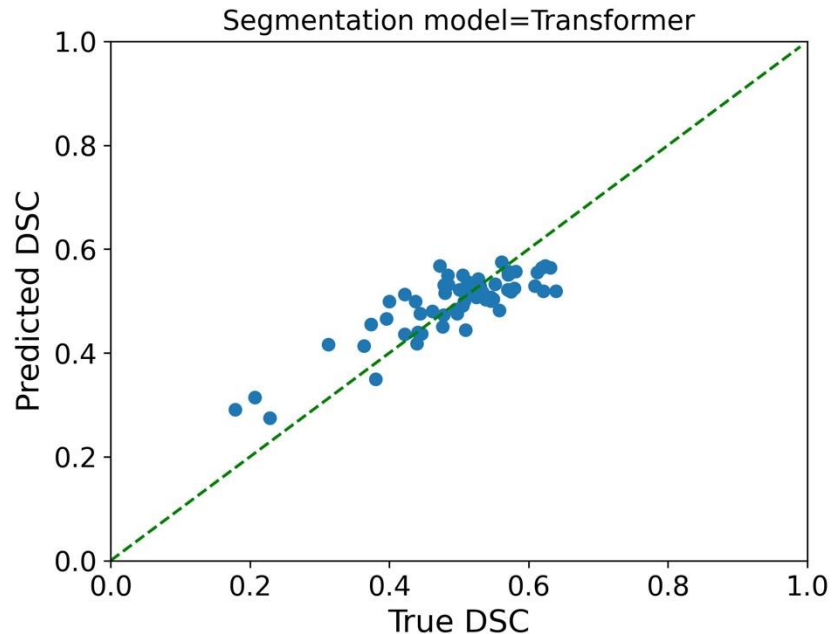
Monitoring and Detection



Results – Regression Model

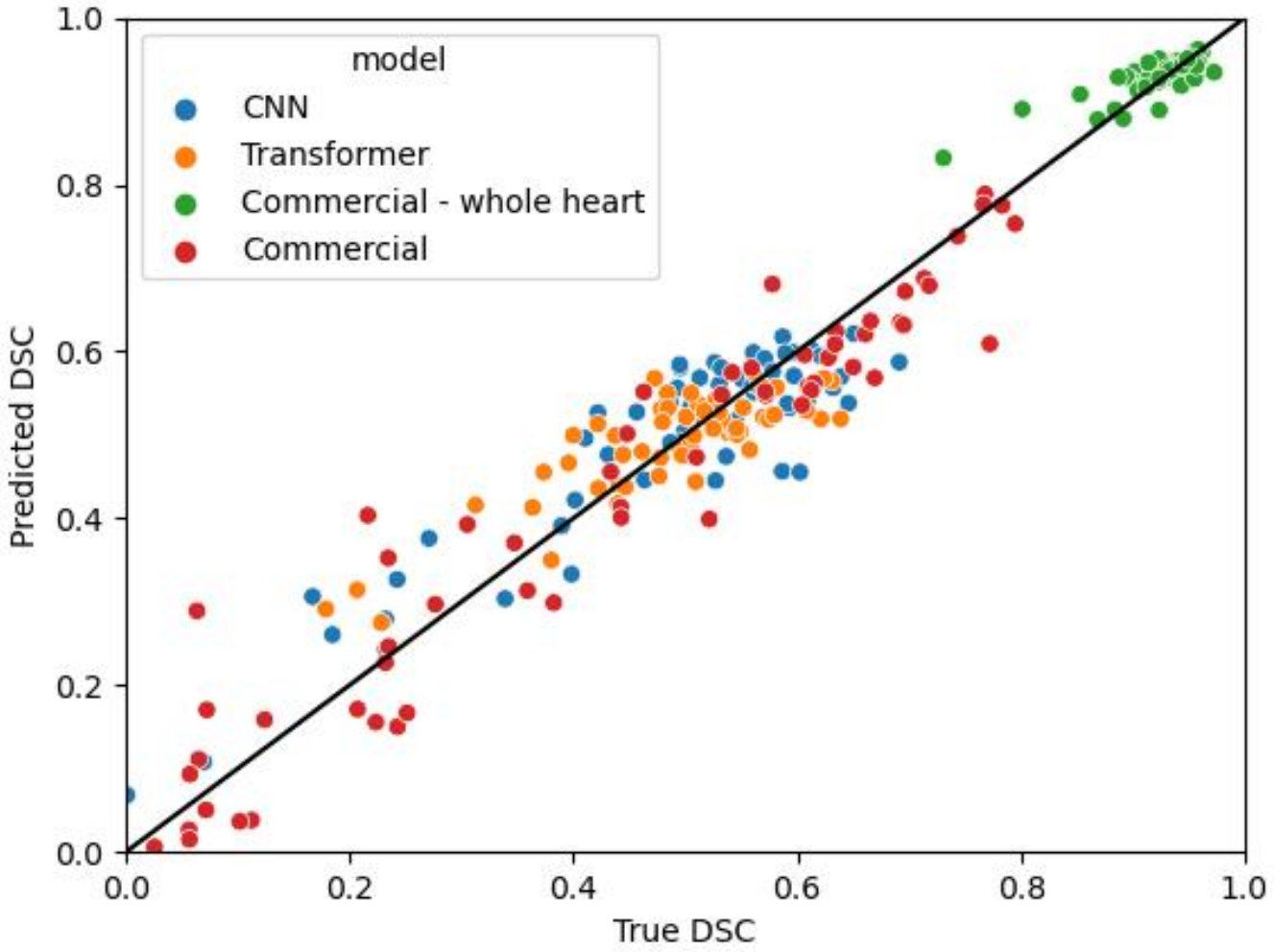
- Autoencoder reconstruction results can then be used to predict autosegmentation system performance
- Able to predict the Dice similarity (autosegmentation vs. physician) within about 10%
- Shows promise for performance monitoring and drift detection



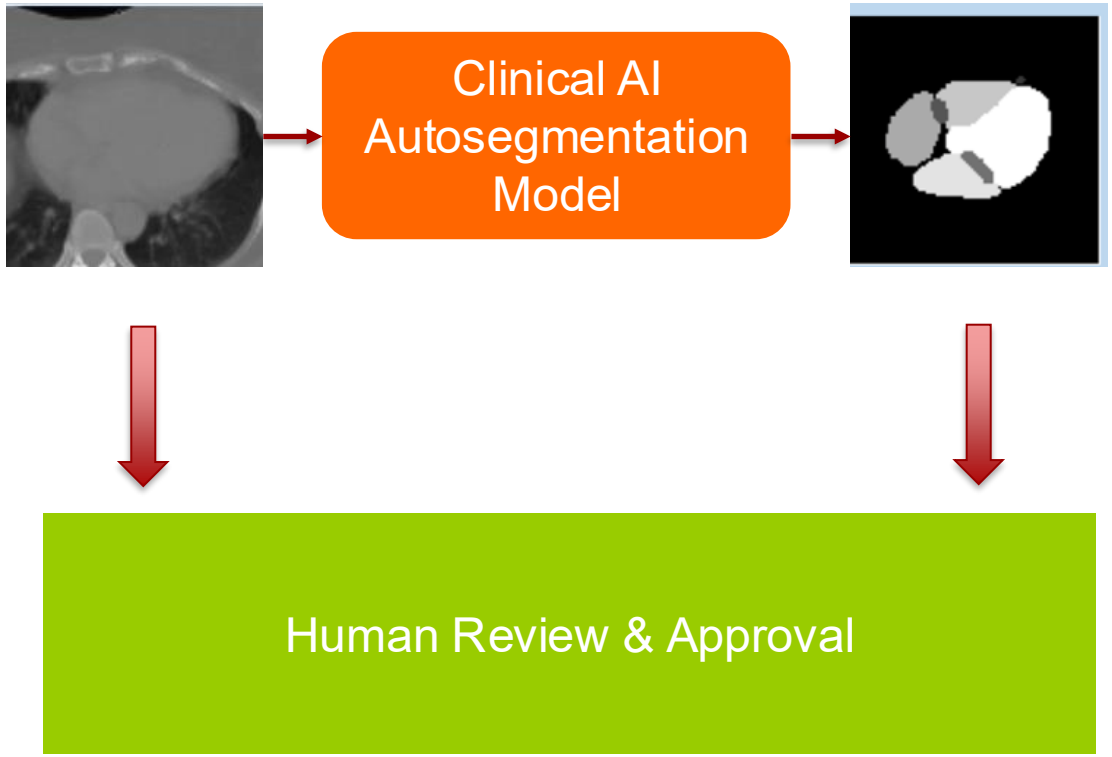


Results – Regression Model

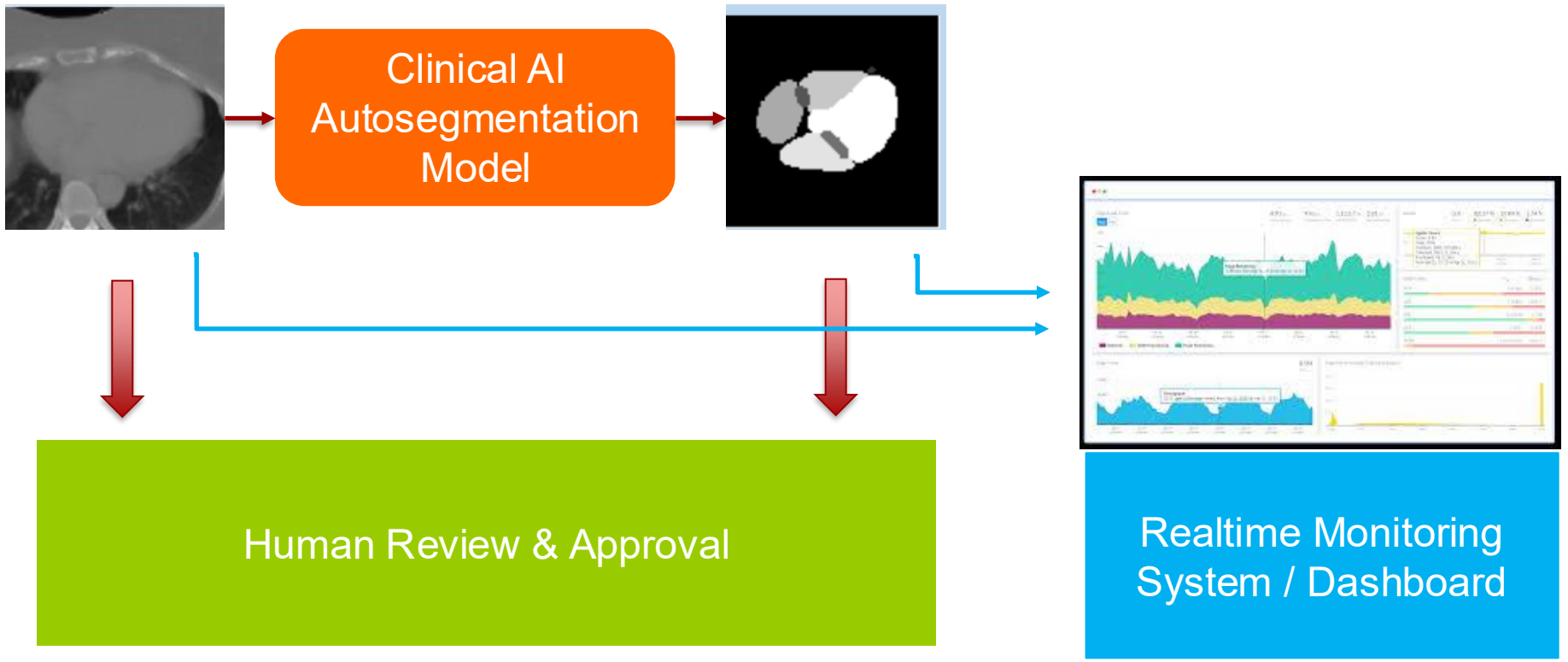
- Autoencoder reconstruction results can then be used to predict autosegmentation system performance
- Able to predict the Dice similarity (autosegmentation vs. physician) within about 10%
- Shows promise for performance monitoring and drift detection



Clinical Deployment – Regression Model

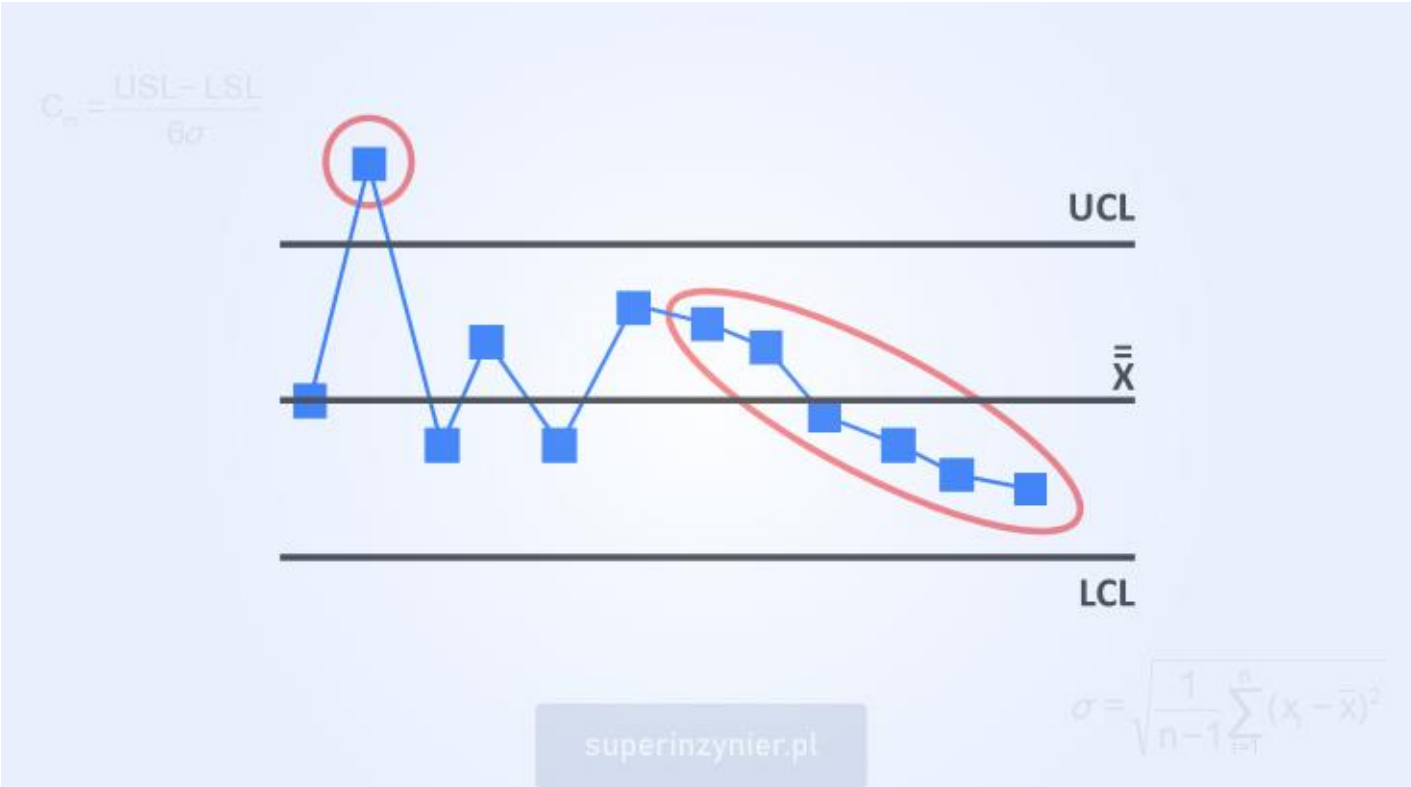


Clinical Deployment – Regression Model



Results – Statistical Process Control

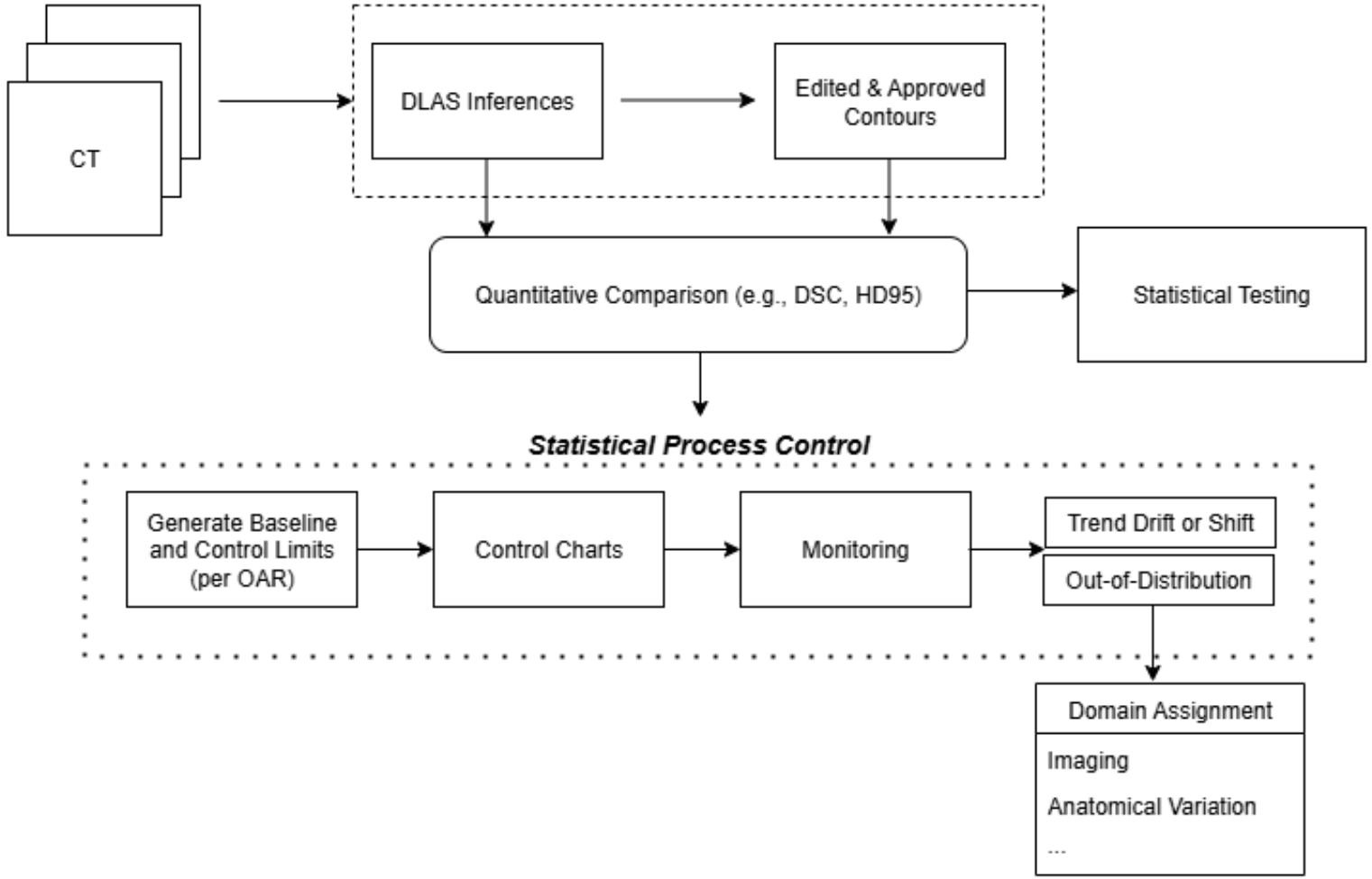
- Statistical process control is a method to monitor and correct a process
- Take a sample, compute a statistic (running mean or median)
- Control chart plots the statistic over time, evaluate deviation and trends



https://www.superengineer.net/blog/spc-introduction

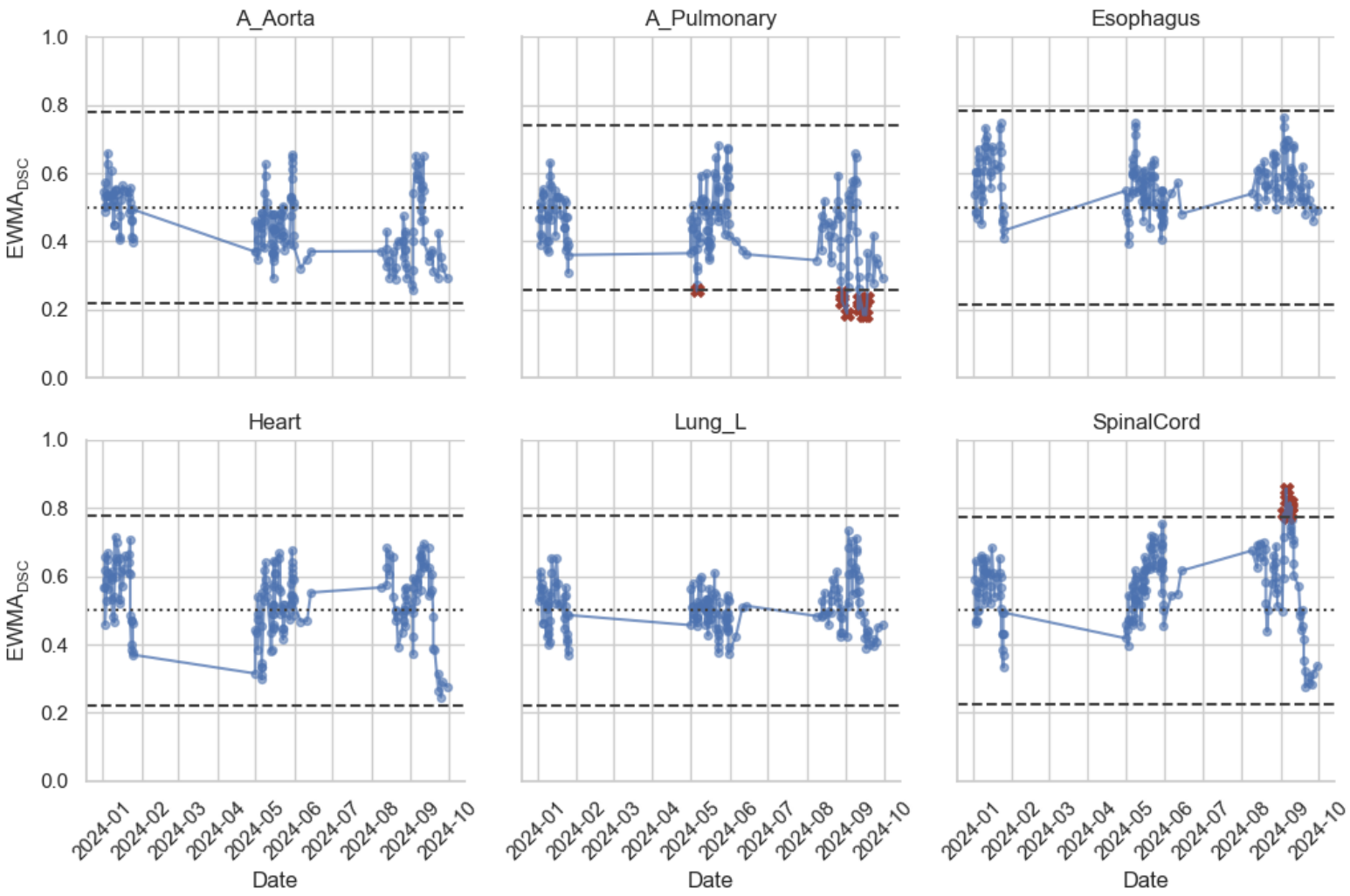
Results – Statistical Process Control

- Clinical autocontouring system evaluated at 4 discrete time points (every 6 months)
- 1st time point used as calibration data
- 295 datasets, 16 organs in thorax and pelvis



Results – Statistical Process Control

- Pulmonary artery and spinal cord contours showed drift out of control range during the 3rd time point
- Other organ contours stable over time

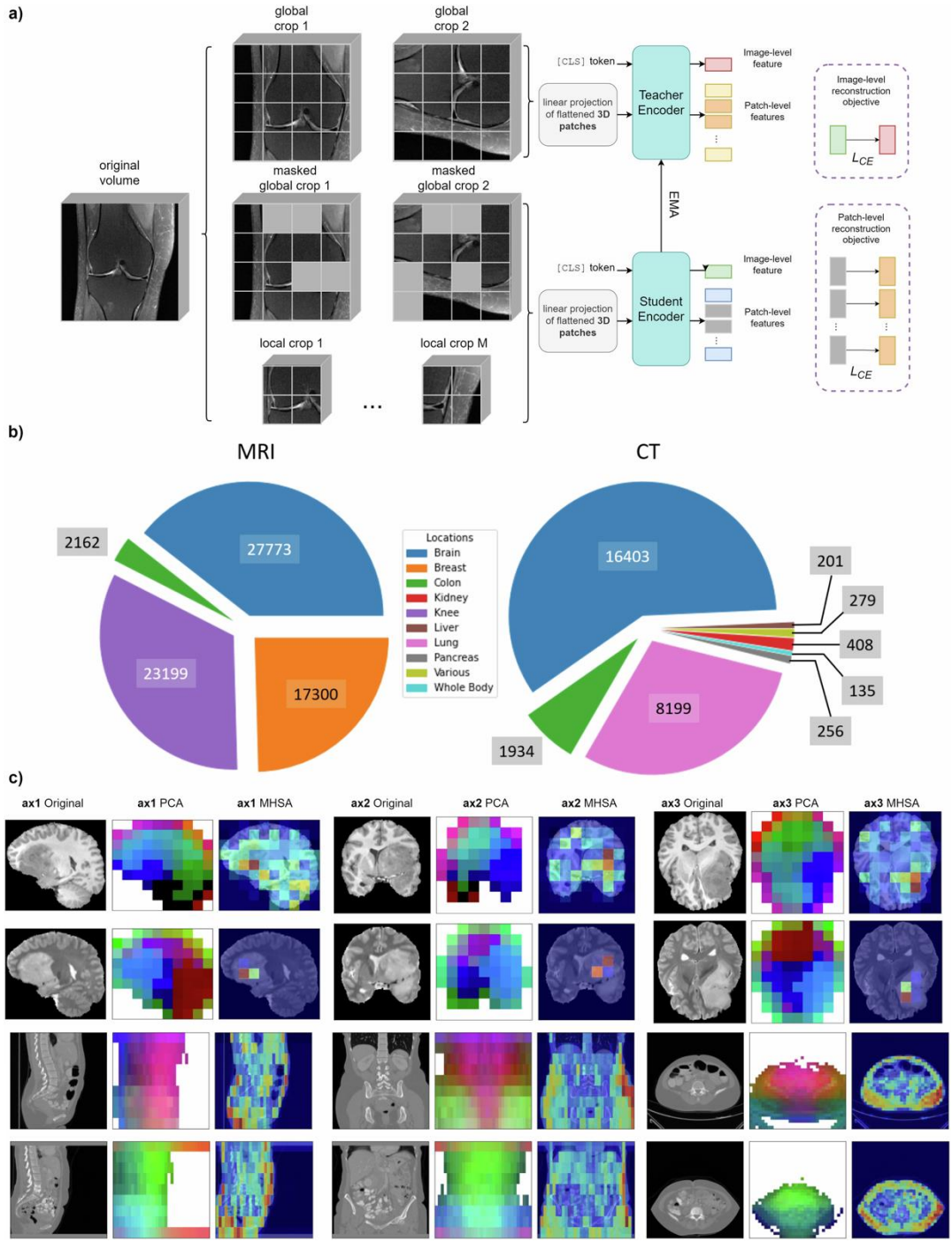


Results – Regression Model

- Challenges
  - Detectors/estimators may need re-training
  - Autoencoders have known deficiencies in self-supervised learning (resolution, representation, etc.)

3DINO for shape autoencoder

- Pretrained Encoder: 3DINO
- Designed for CT, MRI, PET image representation for downstream tasks such as segmentation and classification
- Explicitly developed to produce broadly reusable representations
- Uses Vision Transformer (ViT) which can represent global shape relationships
- Binary contour is still 3D volume, and can be directly inserted to the 3DINO encoder to get latent representation
- ~300M parameters

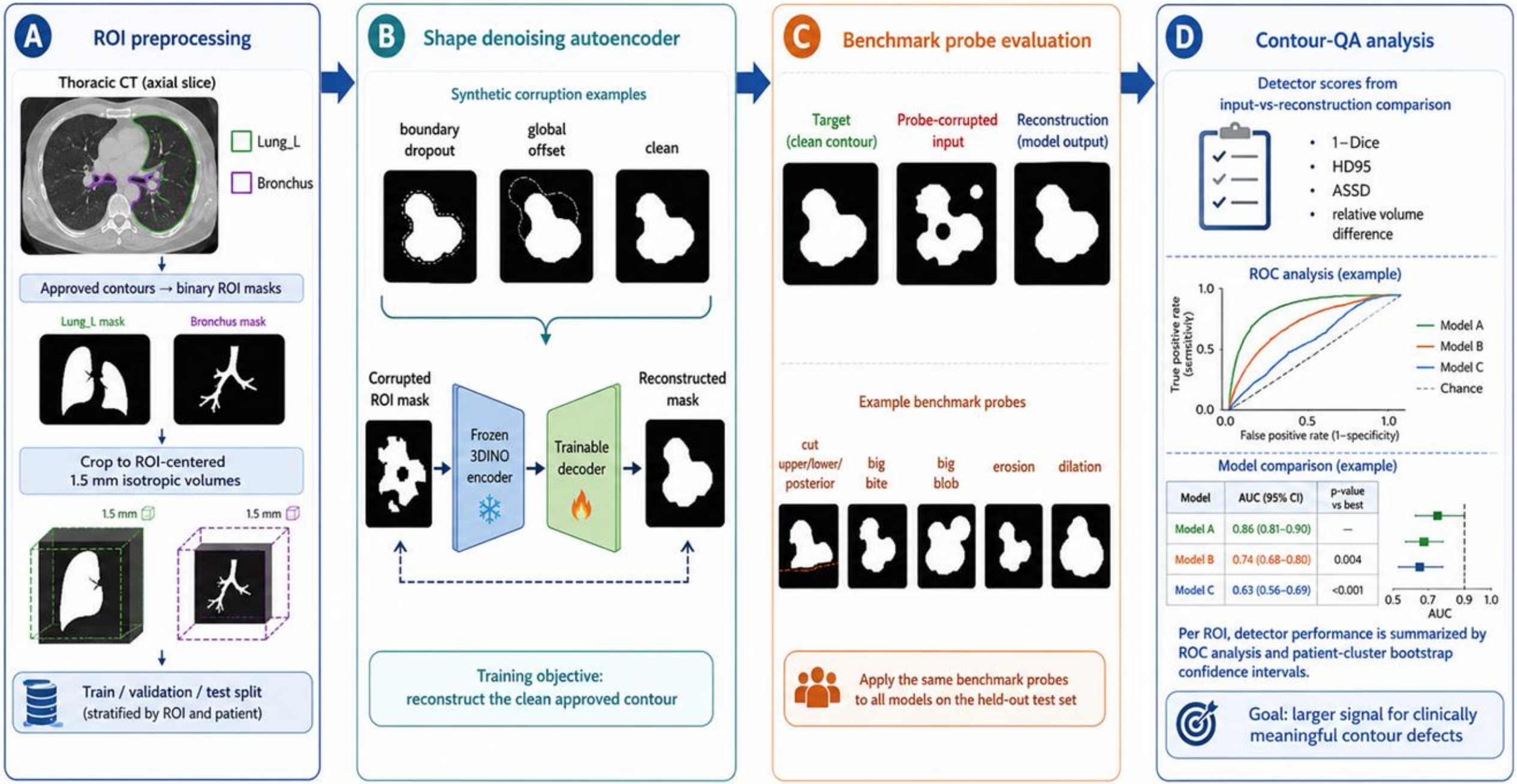


A generalizable 3D framework and model for self-supervised learning in medical imaging

Tony Xu, Sepehr Hosseini, Chris Anderson, Anthony Rinaldi, Rahul G. Krishnan, Anne L. Martel & Maged Goubran

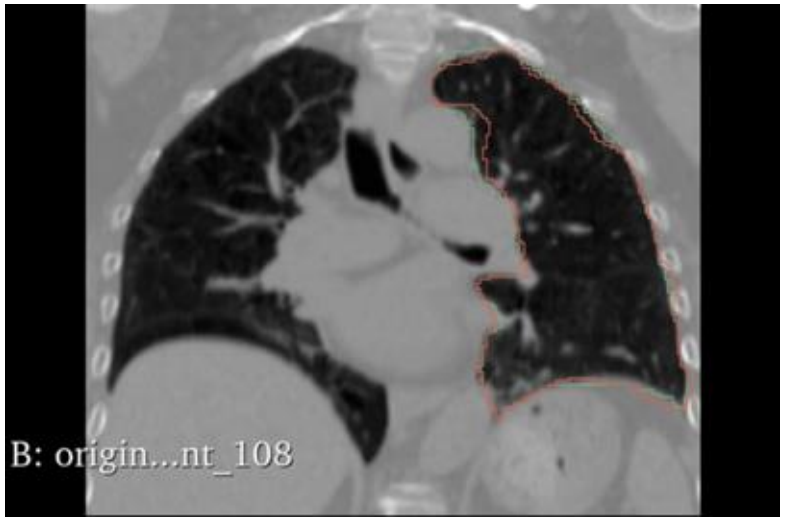
npj Digital Medicine **8**, Article number: 639 (2025) | Cite this article

Figure 1. Overview of the contour-QA framework



Training deformations for shape denoising autoencoder

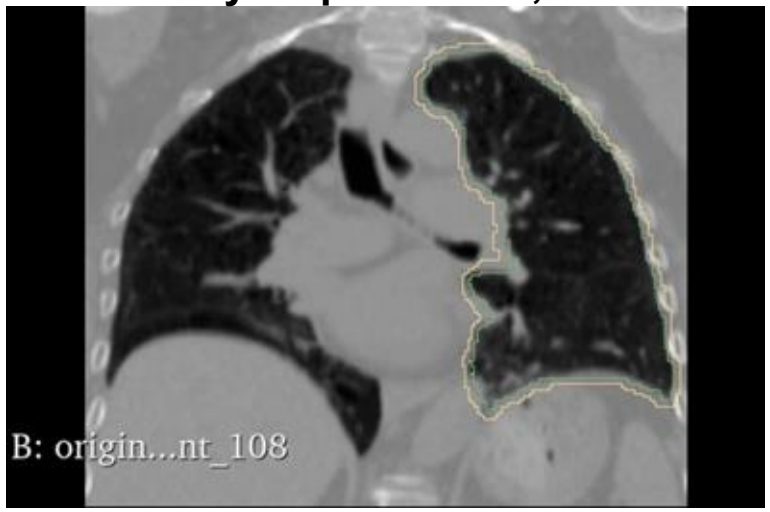
Signed distance perturbation



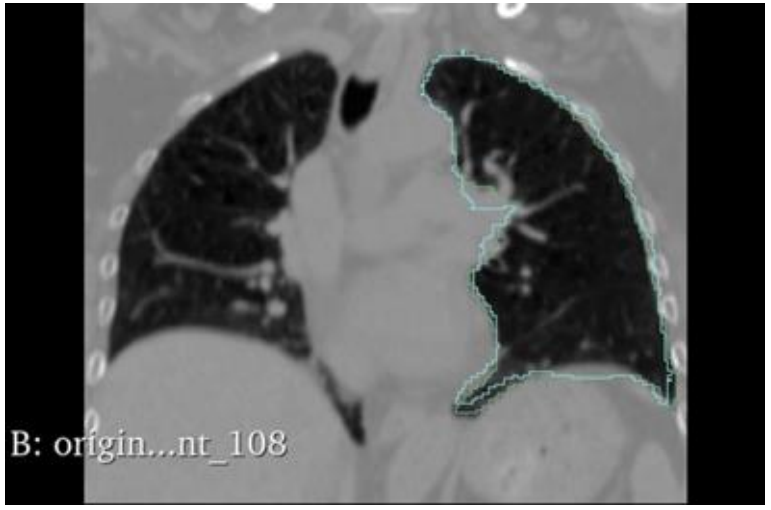
inward global signed-distance boundary displacement



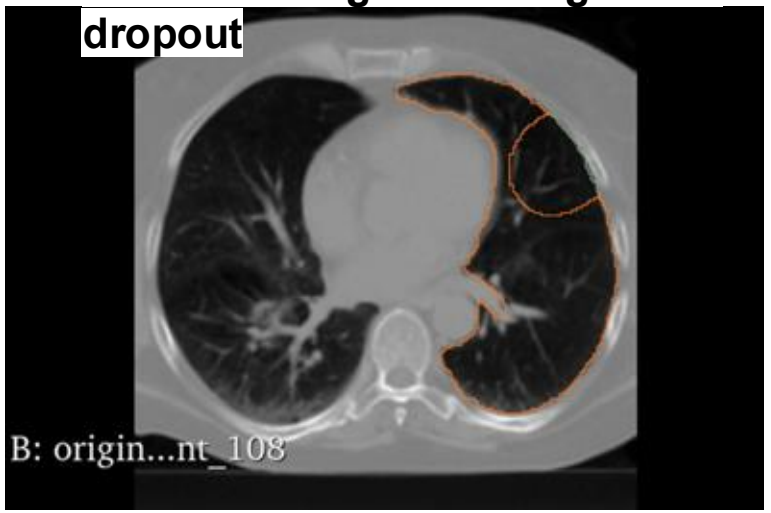
outward global signed-distance boundary displacement,



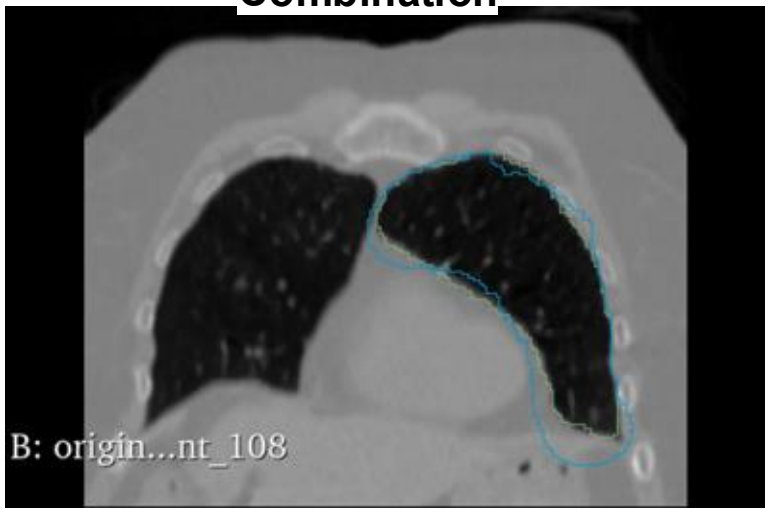
smooth spatial deformation



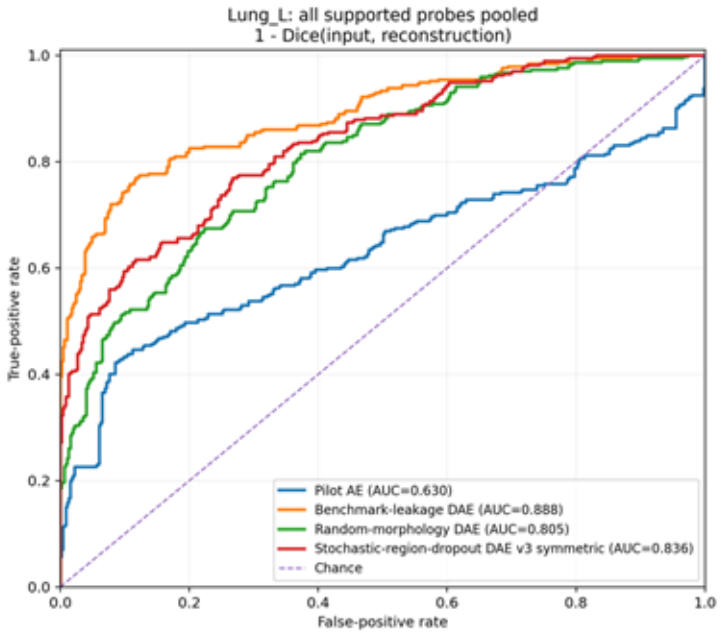
stochastic regional foreground dropout

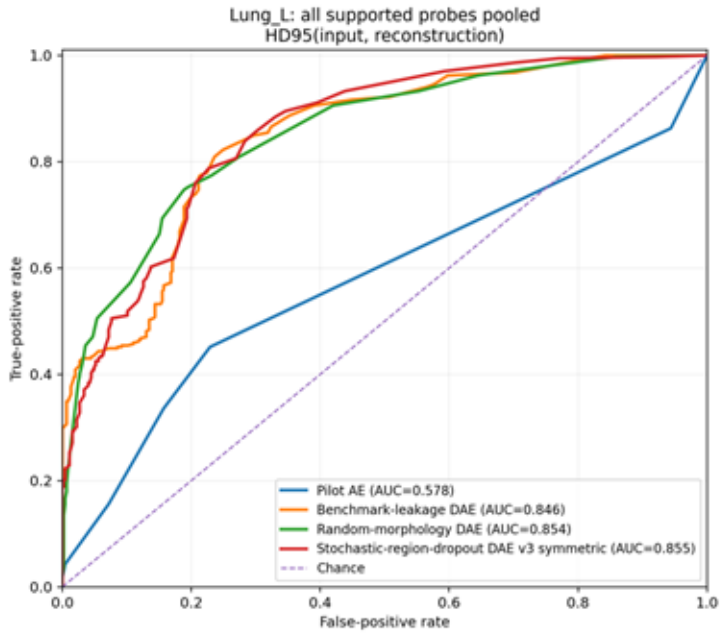


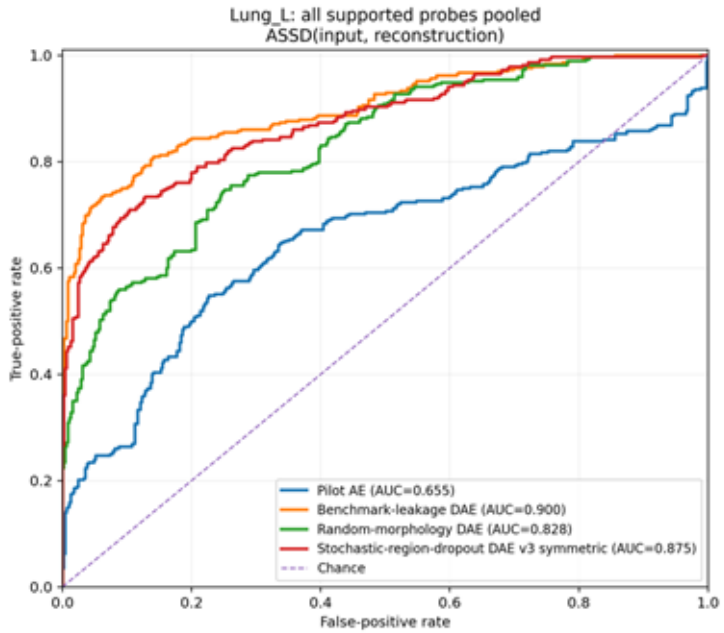
Combination



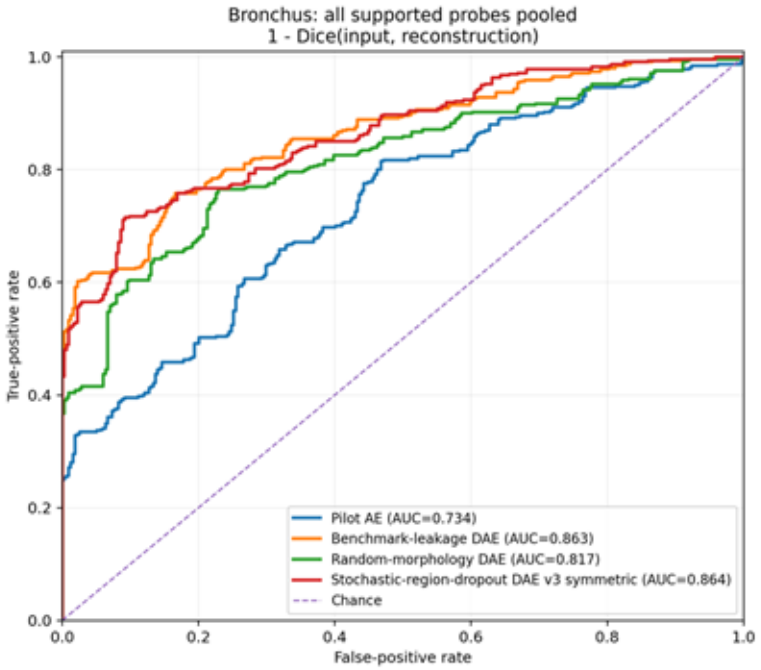
Pooled detector performance: Lung_L

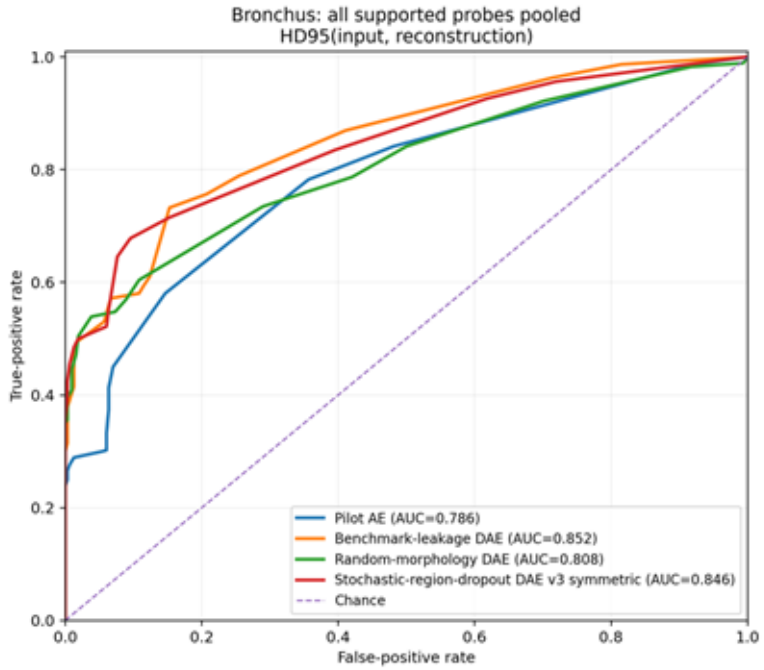


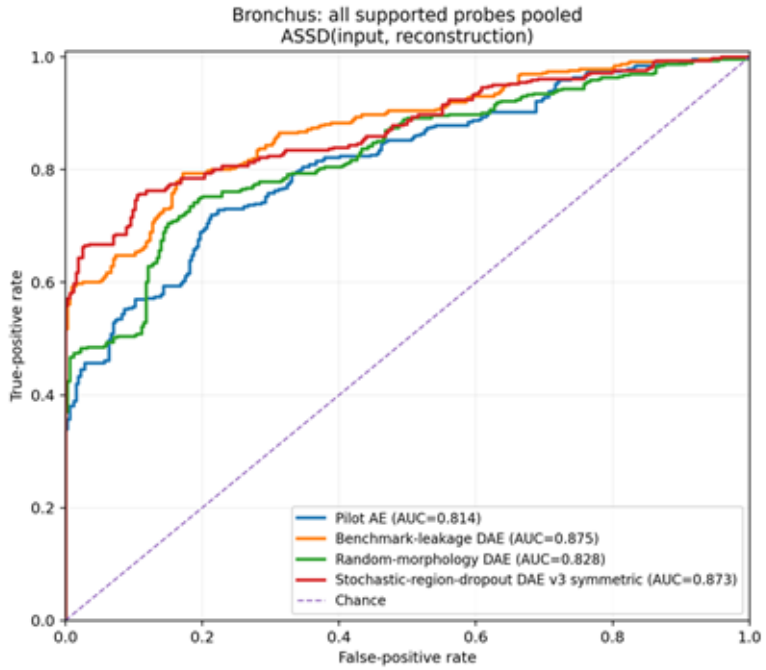




Probe detector performance: Bronchus



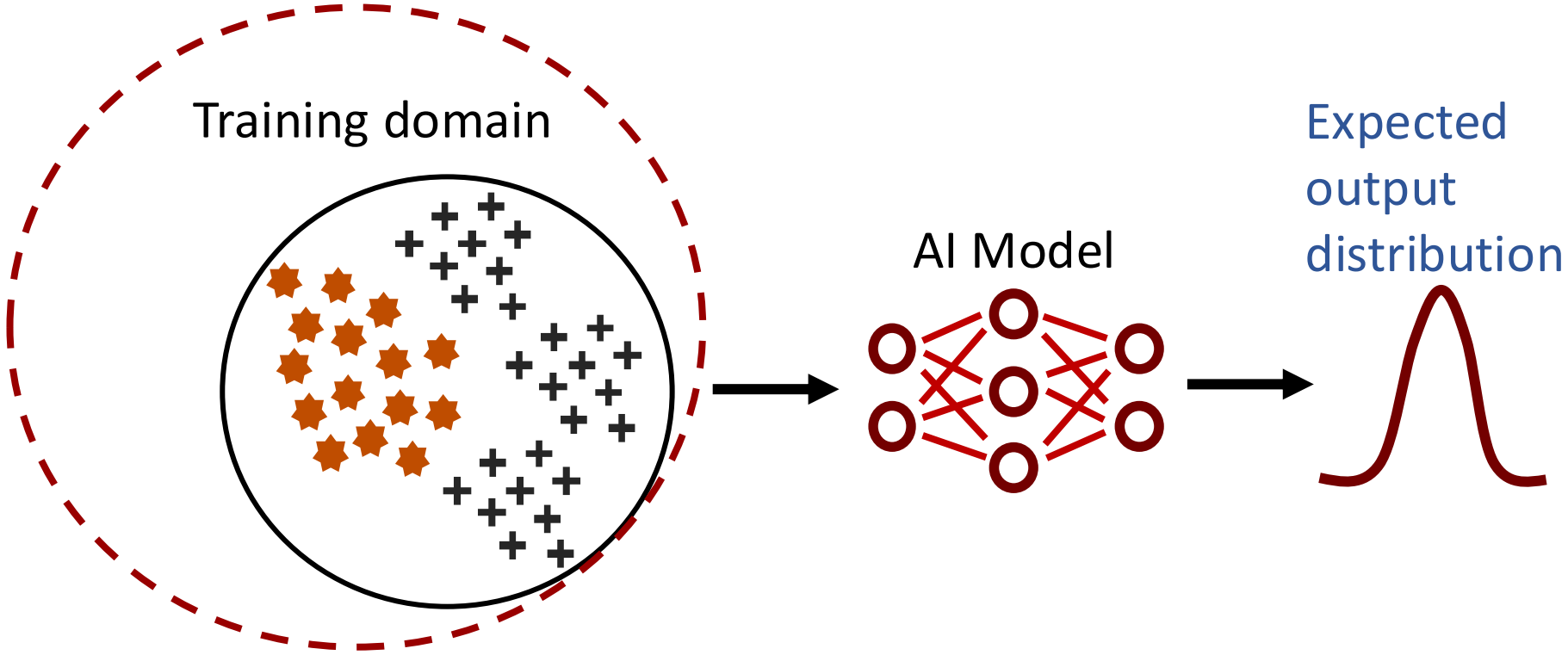




Domain Shift - Solutions

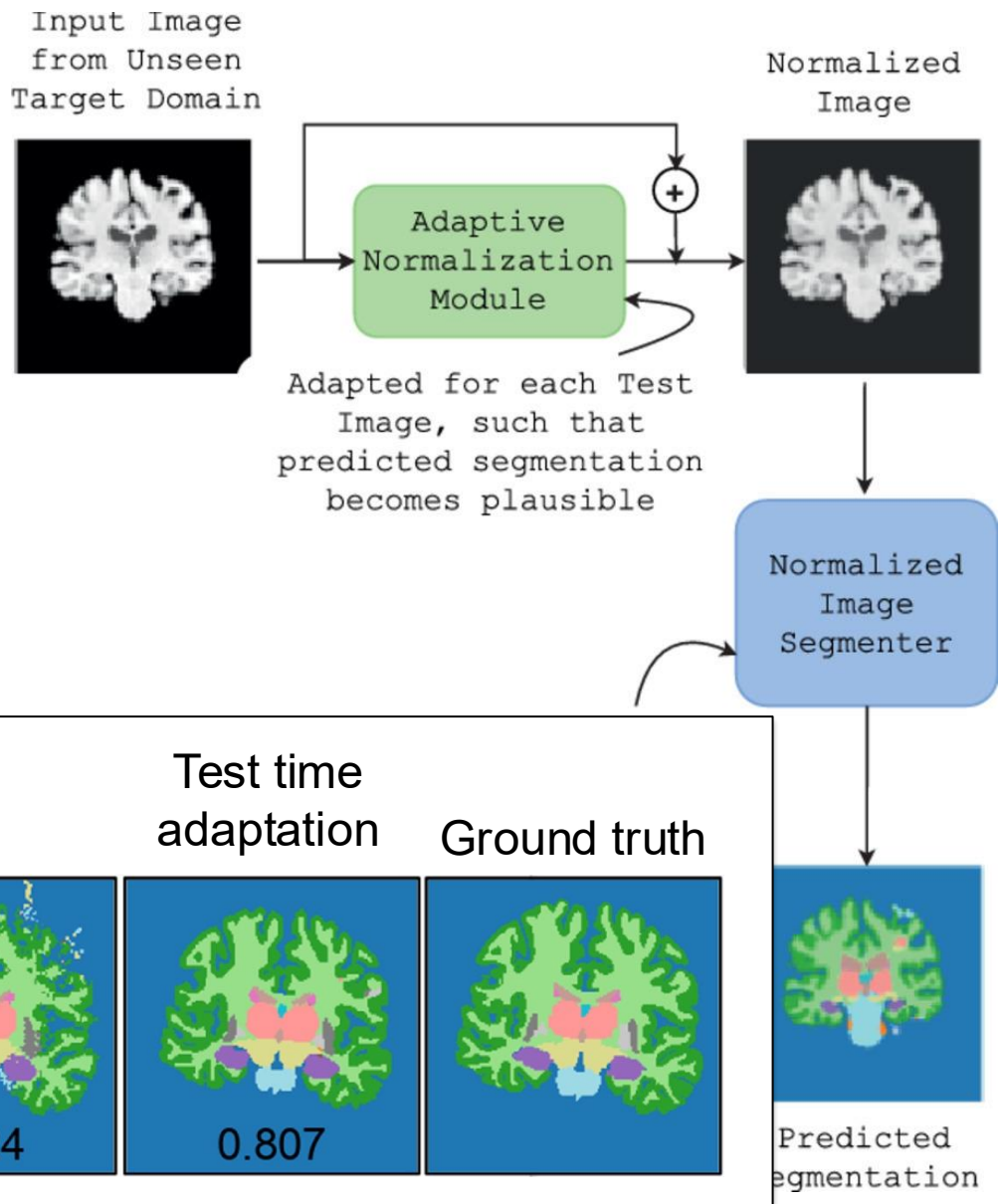
Domain Shift - Solutions

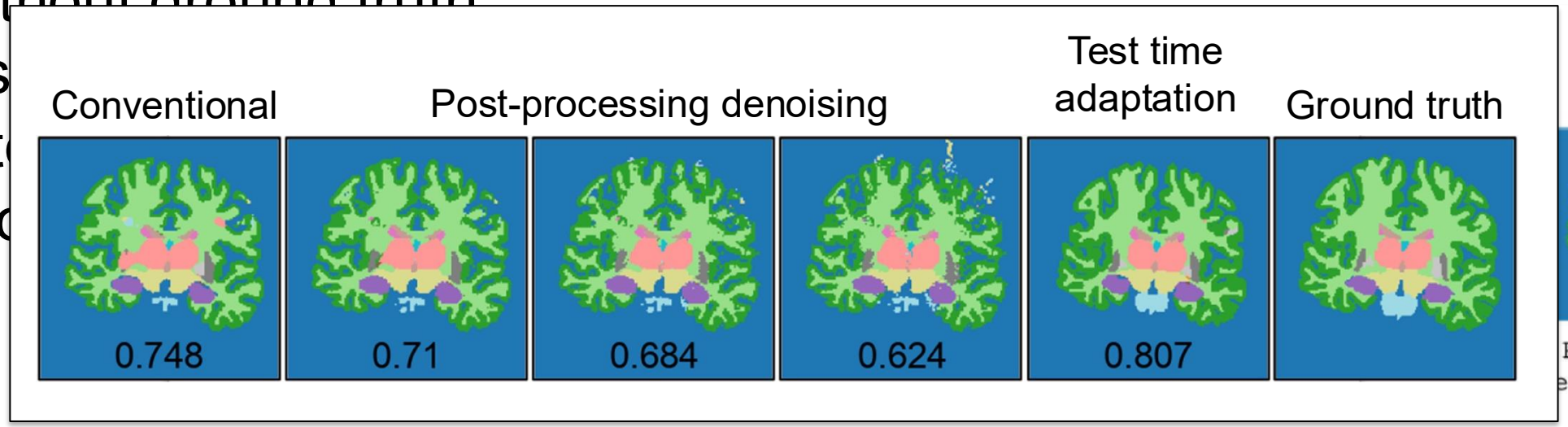
- Add more and varied data from various sources during model training.
  - Synthetic data (augmentation), Collect a huge database from many hospitals
- Problems:
  - requires massive amounts of data (cost, time)
  - May never ‘cover’ the whole problem



Test Time Adaptation

- Add robustness to model training
- Karani et al., Med Imag Anal 2021
  - Concatenate image normalization network and segmentation network
  - Update normalization at “test time” without ground truth
  - As
    int
    pr





https://doi.org/10.1016/j.media.2020.101907

Key Questions / Challenges

- Clinical translation – testing with medical physicists

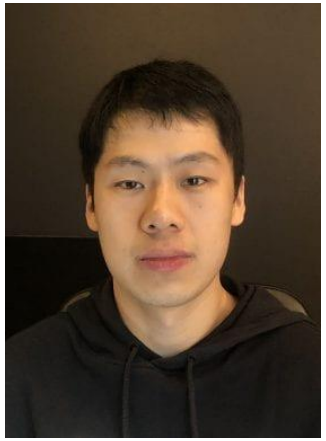
- How to determine *source* of domain shift?
  - Scanner, protocol, clinician, patient, population?
  - For autosegmentation, spatial location of errors may be important to identify

- Can we work with vendors to integration domain shift corrections into the autosegmentation algorithms themselves?

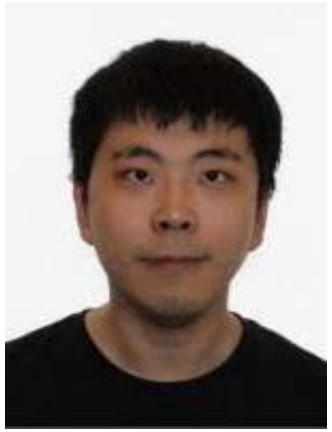
Acknowledgements



Xiyao Jin



Zhehao Zhang



Shaohan Li



Jess Hilliard



Mike Waters



Abhinav Jha



Zhuo Chen



Dean Hobbis

Haydon Windsor (Med Phys PhD)

Received: 30 May 2023 | Revised: 24 October 2023 | Accepted: 30 October 2023
DOI: 10.1002/mp.16846

RESEARCH ARTICLE

MEDICAL PHYSICS

A quality assurance framework for routine monitoring of deep learning cardiac substructure computed tomography segmentation models in radiotherapy

Xiyao Jin[1] | Yao Hao[1] | Jessica Hilliard[1] | Zhehao Zhang[1] | Maria A. Thomas[1] | Hua Li[1] | Abhinav K. Jha[2,3] | Geoffrey D. Hugo[1]

Received: 17 December 2021 | Revised: 19 July 2021 | Accepted: 13 September 2021
DOI: 10.1002/mp.15237

RESEARCH ARTICLE

MEDICAL PHYSICS

Robustness of deep learning segmentation of cardiac substructures in noncontrast computed tomography for breast cancer radiotherapy

Xiyao Jin | Maria A. Thomas | Joseph Dise | James Kavanaugh | Jessica Hilliard | Imran Zoberi | Clifford G. Robinson | Geoffrey D. Hugo

Funding: NIH R01CA289884, R01CA258821

Professional Training

- Academic Training
  - MS or PhD in medical physics, or
  - MS or PhD in physics or related discipline with post-graduate academic training in medical physics - Graduate Certificate
- Clinical Training – required for board certification
  - Residency in clinical medical physics

More information….

- Wash U Radiation Oncology (radonc.wustl.edu)
- Hugo Lab (hugolab.wustl.edu)
- American Association of Physicists in Medicine (www.aapm.org)
- American Society for Therapeutic Radiology and Oncology (www.astro.org)
- National Cancer Institute (www.cancer.gov)