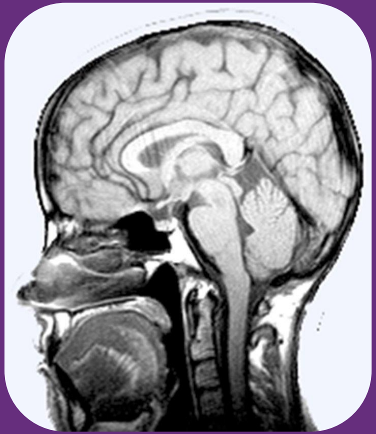
Building Trustworthy Medical AI Systems for Safe Clinical Deployment

Créer des systèmes d'IA médicale fiables pour un déploiement clinique sécurisé



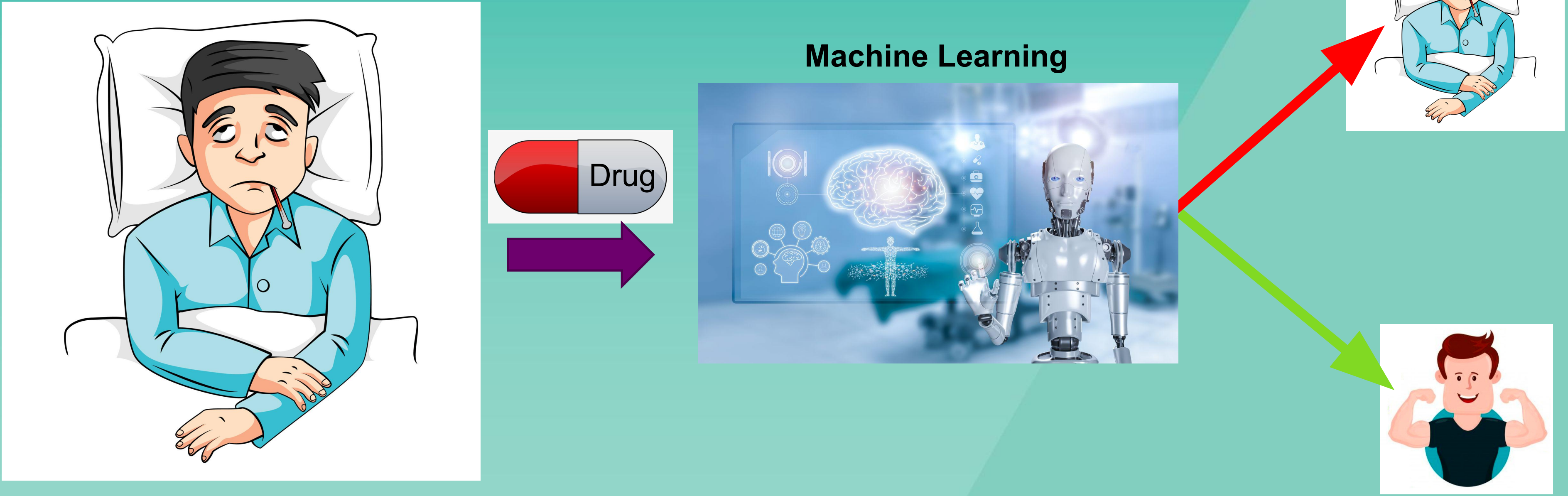
Tal Arbel, PhD

Canada CIFAR AI Chair, Mila
Professor, McGill University, Department of Electrical and Computer Engineering
Director Probabilistic Vision Group, Medical Imaging Lab
Centre for Intelligent Machines

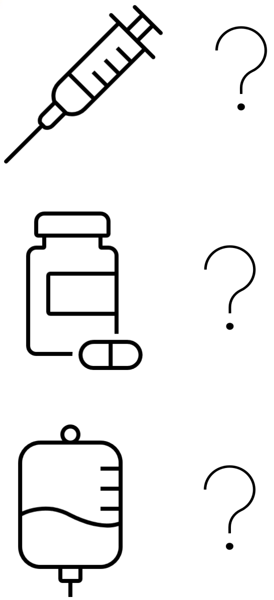
McGill UNIVERSITY CIM CENTRE FOR INTELLIGENT MACHINES PVG Mila

AI for Personalized Medicine:

The Dream and the Challenges

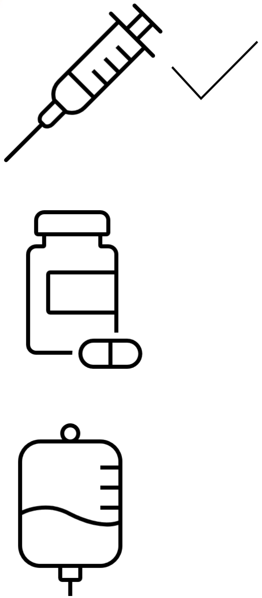


Clinical Scenario - Current Practice



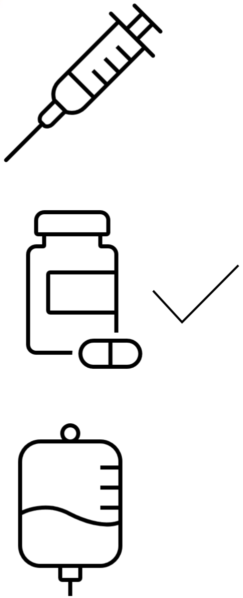
- Variety of treatments available for this patient’s illness

Clinical Scenario - Current Practice



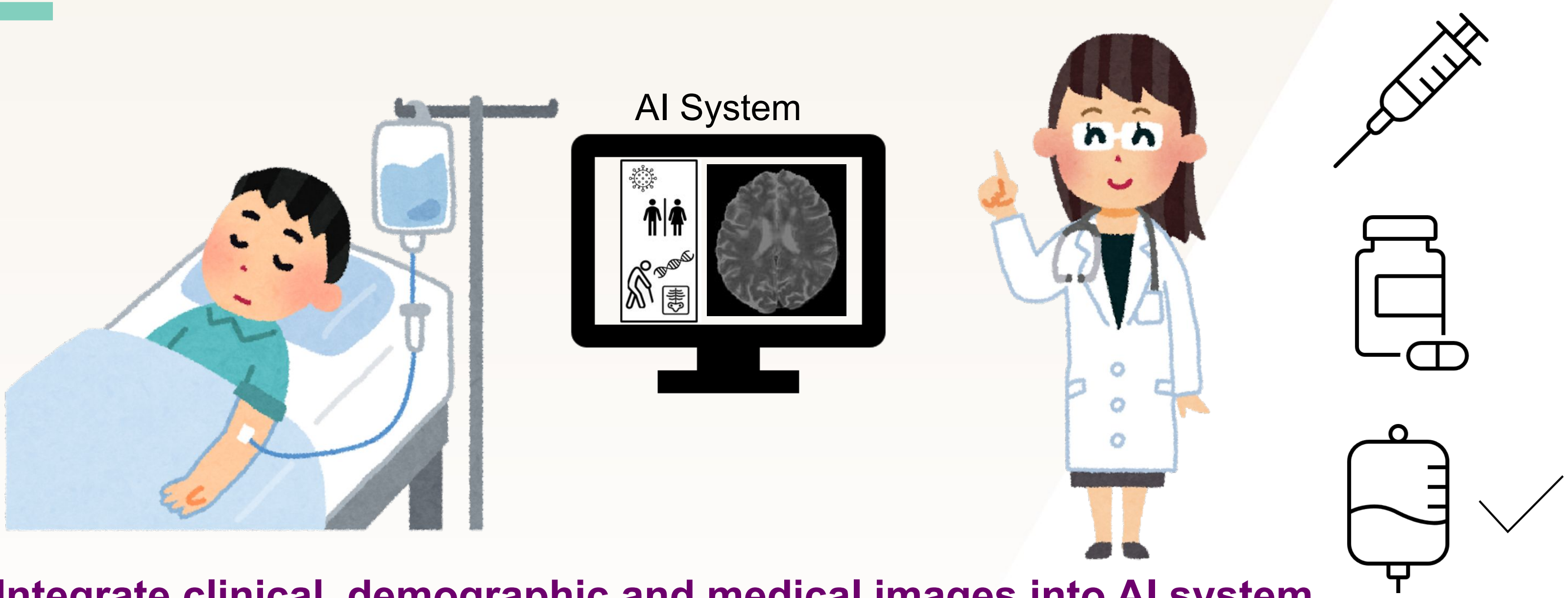
- Variety of treatments available for this patient’s illness
- **Treatment decision:** Average treatment efficacy across population

Clinical Scenario – Personalized Medicine



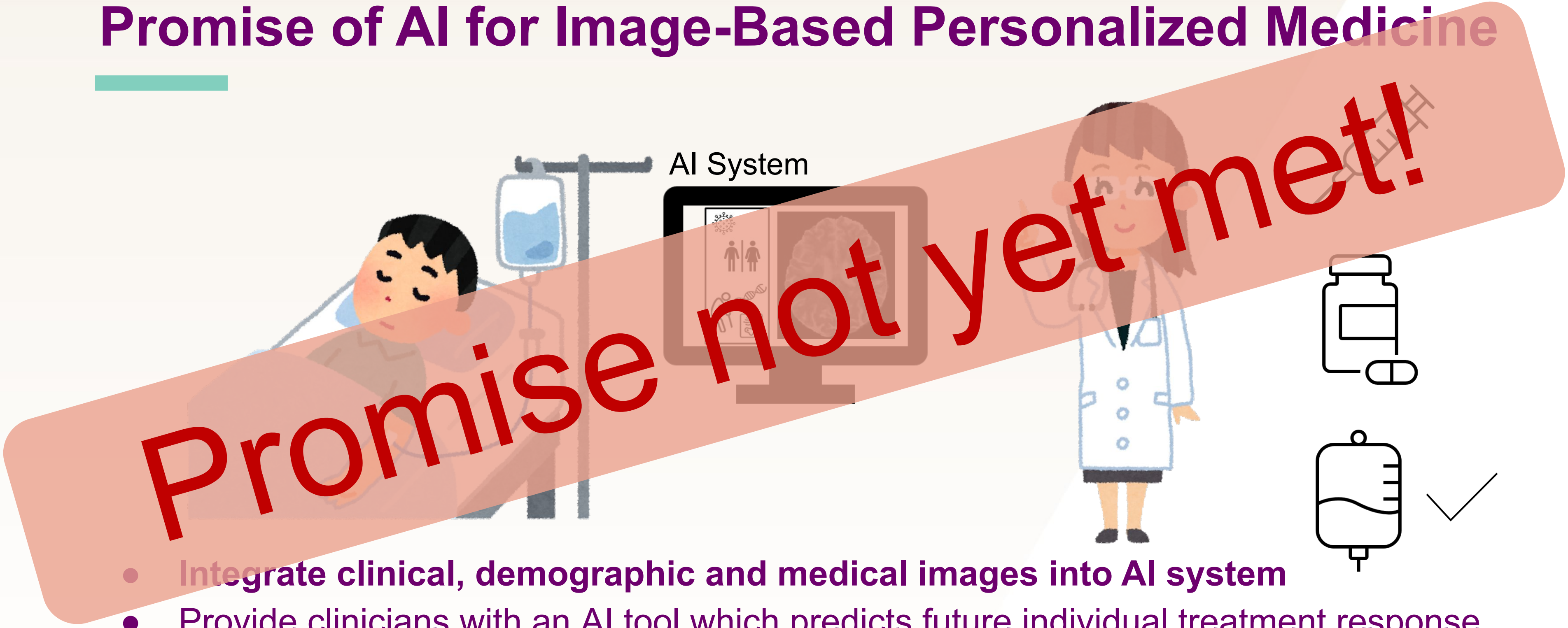
- Clinical and demographic information available
- **Treatment decision:** Average treatment efficacy conditioned on sub-group statistics

Promise of AI for Image-Based Personalized Medicine



- **Integrate clinical, demographic and medical images into AI system**
- Provide clinicians with an AI tool which predicts future individual treatment response on several treatments using *discovered* image features

Promise of AI for Image-Based Personalized Medicine



- **Integrate clinical, demographic and medical images into AI system**
- Provide clinicians with an <u>AI tool</u> which predicts future individual treatment response on several treatments using <u>*discovered* image features</u>

McGill UNIVERSITY CIM CENTRE FOR INTELLIGENT MACHINES PVG Mila

Deep Learning Models Can Make (Potentially Deadly) Mistakes

ARTIFICIAL INTELLIGENCE

Hundreds of AI tools have been built to catch covid. None of them helped.

Commentary

Trustworthy medical AI systems need to know when they don't know

Thomas Grote

Correspondence to Thomas Grote, Ethics and Philosophy Lab, University of Tübingen, Tübingen 72076, Germany; thomas.grote@uni-tuebingen.de

https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

Lack of Interpretability of Deep Learning Models

This patient has pleural effusion

How did the classifier come to this conclusion?

Need to open up the black box!

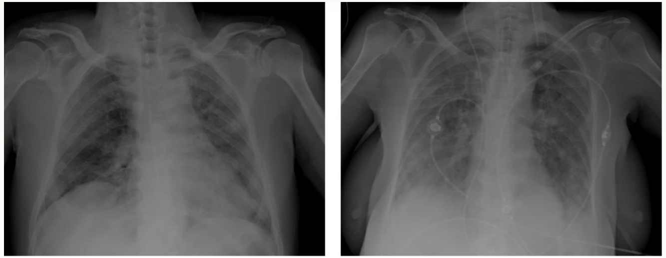
Deep Learning Models Can Be Biased

BRIEF REPORT | APPLIED MATHEMATICS

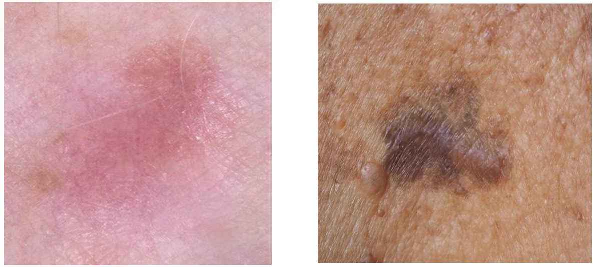
Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

AI skin cancer diagnoses risk being less accurate for dark skin - study

Research finds few image databases available to develop technology contain details on ethnicity or skin type



(a) Male (b) Female



→ Need to mitigate the biases

https://www.pnas.org/doi/10.1073/pnas.1919012117
https://www.theguardian.com/society/2021/nov/09/ai-skin-cancer-diagnoses-risk-being-less-accurate-for-dark-skin-study

Talk

(1) **First** deep learning model for personalized medicine from patient images

(2) Trustworthiness and reliability of deep learning models needed in clinical applications:

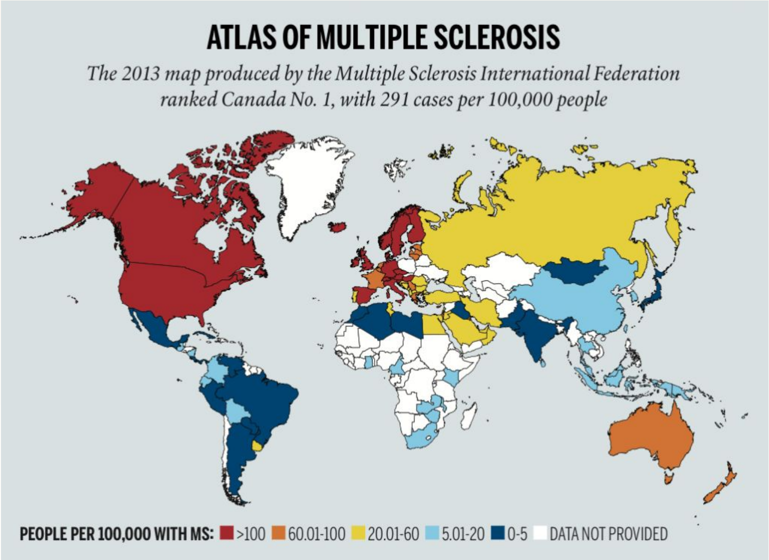
- Uncertainty Estimation
- Explainability
- Improving fairness/biases

Vision-Language Foundation Models

Case studies: Multiple Sclerosis MRI; Chest Xrays

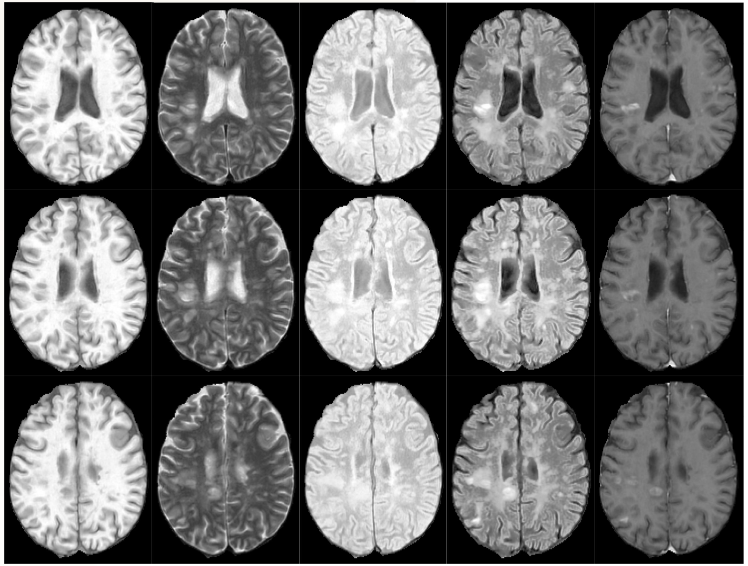
Case Study: Multiple Sclerosis

Most common neurological disease affecting young people; Canada has highest rate per capita.



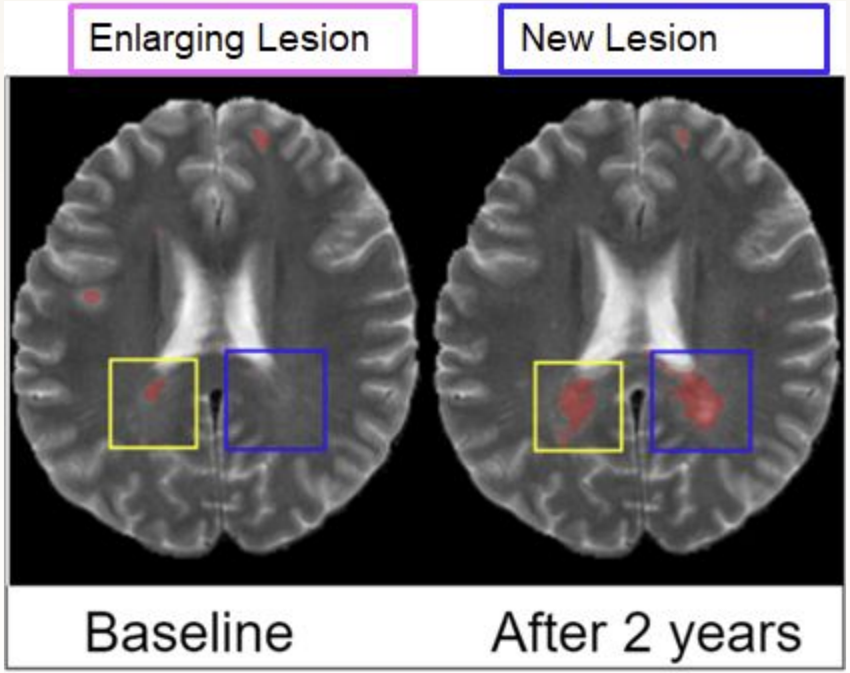
https://macleans.ca/society/health/could-canada-cause-multiple-sclerosis/

Multi-focal brain lesions visible on MRI



T1w T2w PDw FLR T1c

Case Study: Multiple Sclerosis



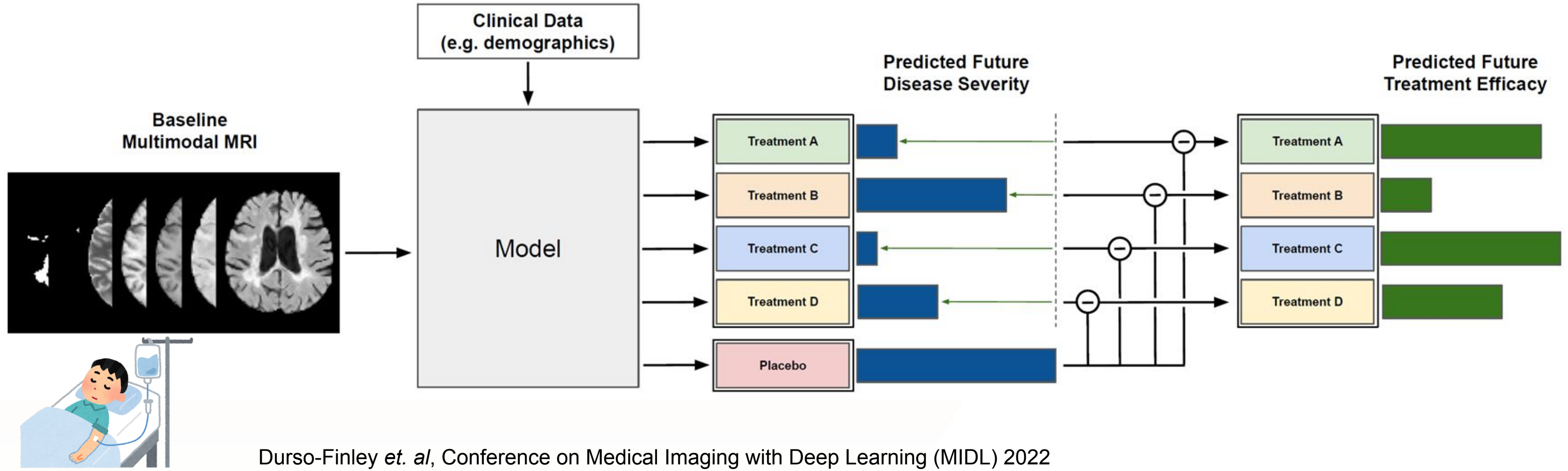
Appearance of new/enlarging lesions on successive MRI scans important:

- MRI markers of new disease activity since previous scan
- Treatments exist to help suppress new lesions, manage symptoms (not to stop progression)
  - Different efficacies
  - Risk profiles, side effects, cost, availability, etc.
- No Cure.

Deep Learning for Image-Based Precision Medicine

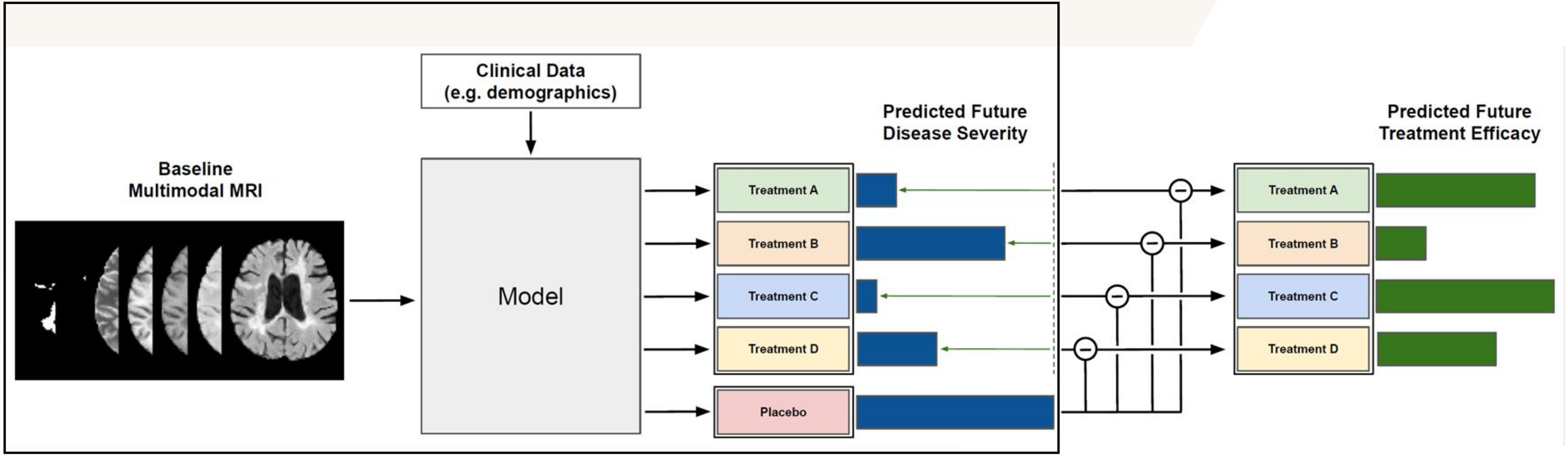
First *DL model that learns data driven imaging markers predictive of future disease progression for individual patients on and off treatment*

"Big Data": Largest dataset of MS patient images acquired during clinical trials



Durso-Finley *et. al*, Conference on Medical Imaging with Deep Learning (MIDL) 2022

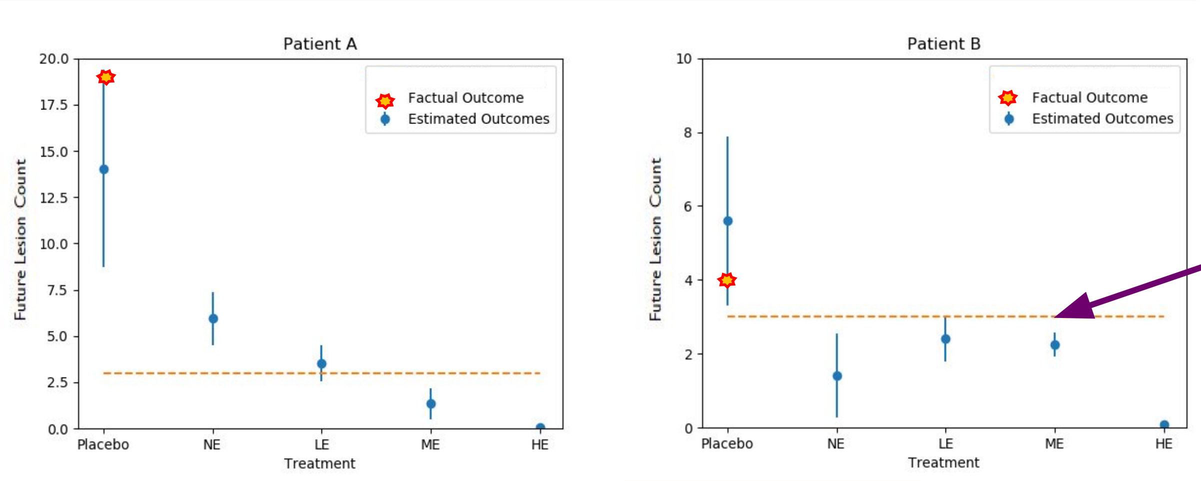
Deep Learning for Personalized Prediction of Future Outcomes on and off Treatment from Images



Durso-Finley *et. al*, Conference on Medical Imaging with Deep Learning (MIDL) 2022

Deep Learning for Clinical Decision Support

Individual Future Treatment Outcome Estimates

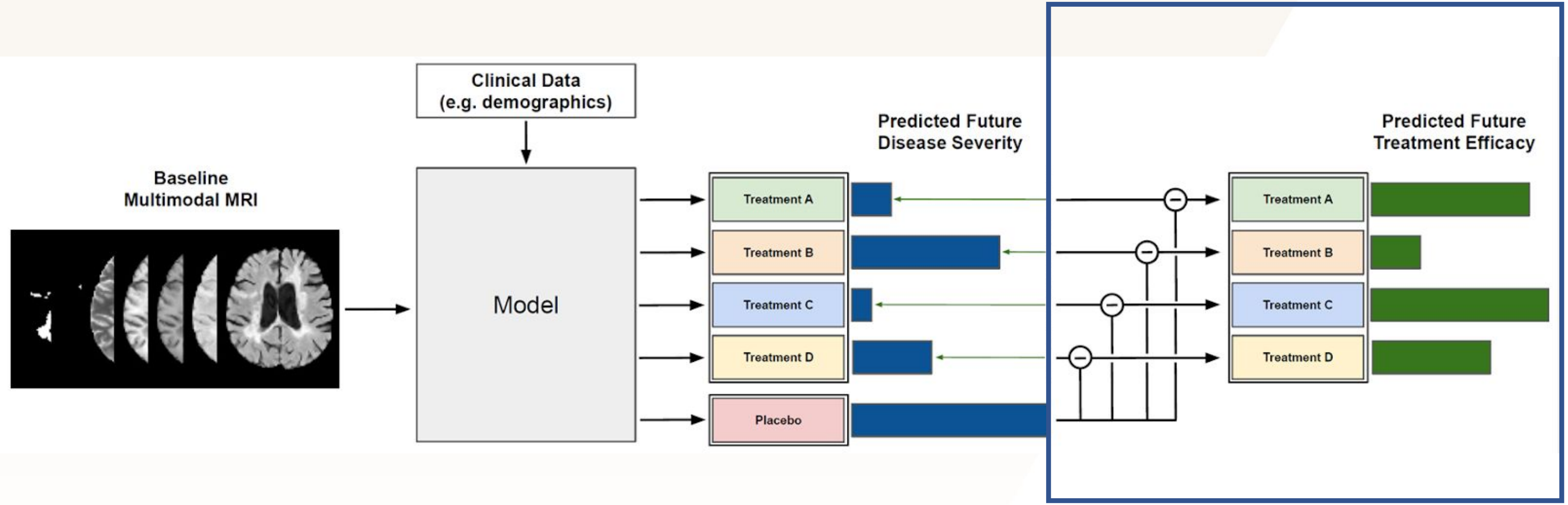


Established drug escalation threshold

NE: No proven efficacy, **LE**: Lesser efficacy , **ME**: Moderate efficacy, **HE**: High efficacy

Deep Learning for Clinical Decision Support

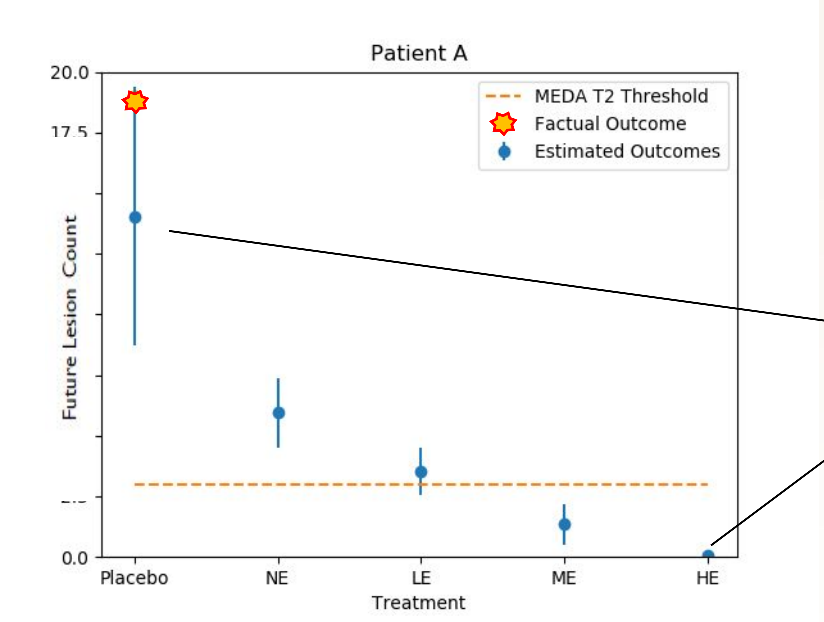
Predicting Future Treatment Effects



→ *Causal effect of treatment on the outcome for a patient*

McGill University CIM Centre for Intelligent Machines PVG Mila

Estimating Future Personalized Treatment Response



Predicted Future Individual Treatment Response

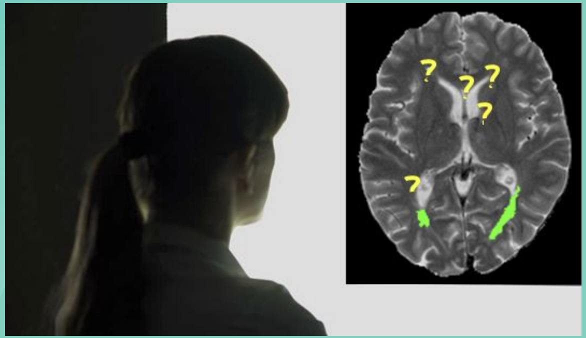
→ Reduction in future new lesion count relative to placebo (no treatment)

NE: No proven efficacy, **LE**: Lesser efficacy , **ME**: Moderate efficacy, **HE**: High efficacy

Are we ready for clinical deployment?

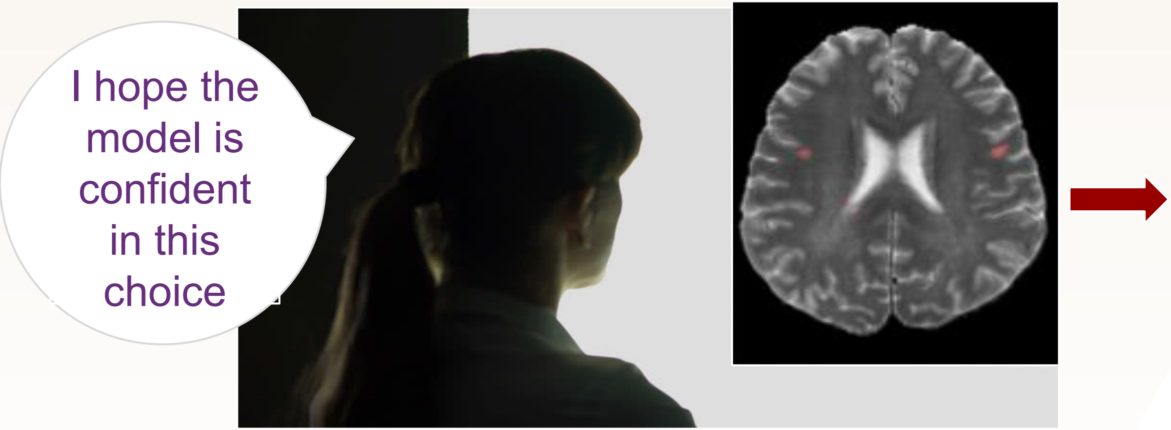
Trustworthy Image-Based Personalized Medicine

- Uncertainty estimation
- Explainability
- Improving biases



Trustworthy Image-Based Personalized Medicine

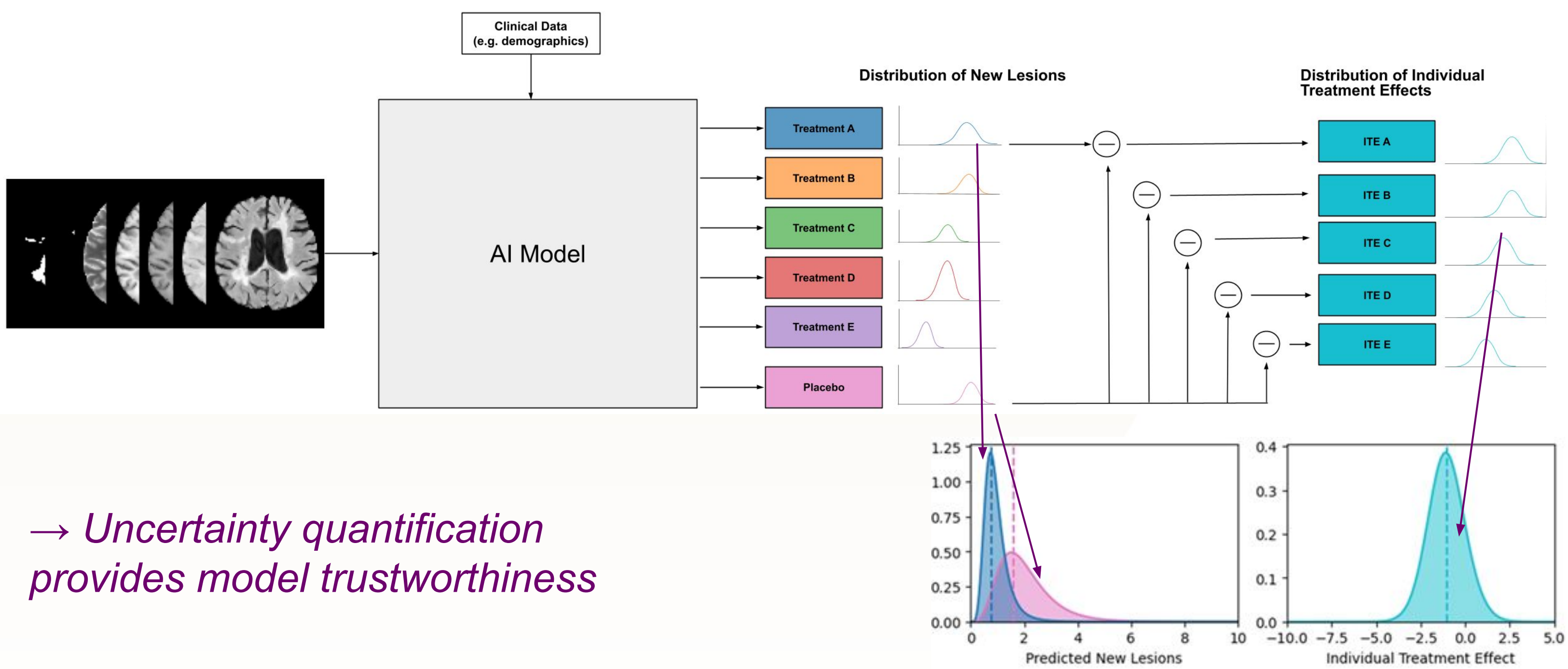
AI makes mistakes! High risk in handing over to clinician



"Patient will be a future responder to Treatment A"

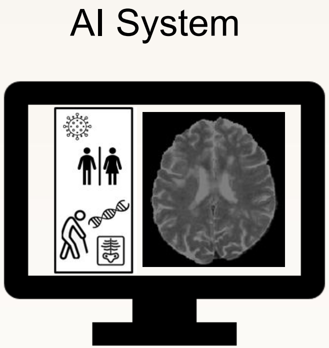
→ What if we could quantify reliability of predictions in the form of uncertainty?

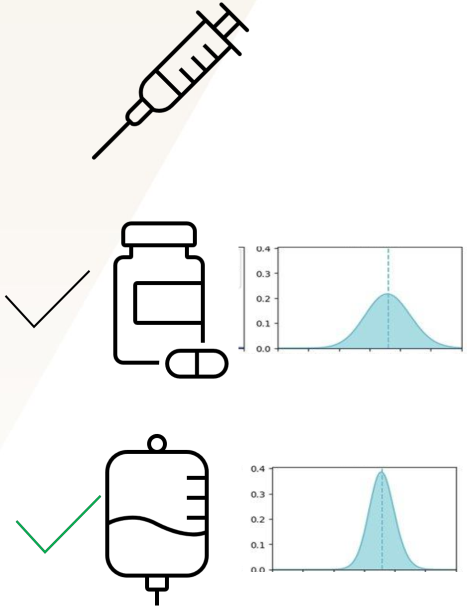
Trustworthy Treatment Effect Estimation



→ Uncertainty quantification provides model trustworthiness

The Promise of AI for Image-Based Personalized Medicine





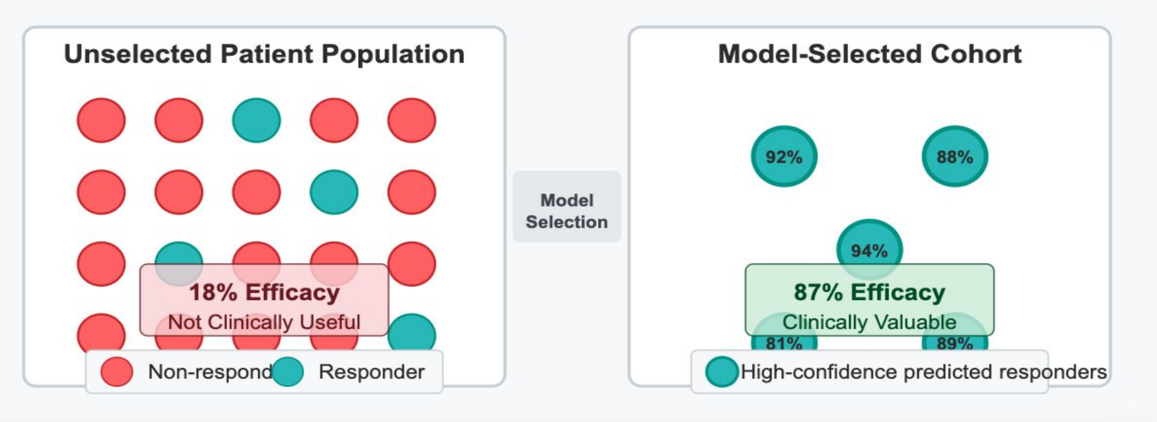
→ AI tool predicts patient will respond to both treatments *but is more confident in last one!*

Finding Responders in MS Clinical Trials

Selecting the patients with higher probabilities of response

→ Model identified responders to drugs with high confidence *even with low efficacy drugs*!

- Patients have **more options in the clinic**
- Finds **good candidates for trial enrichment**

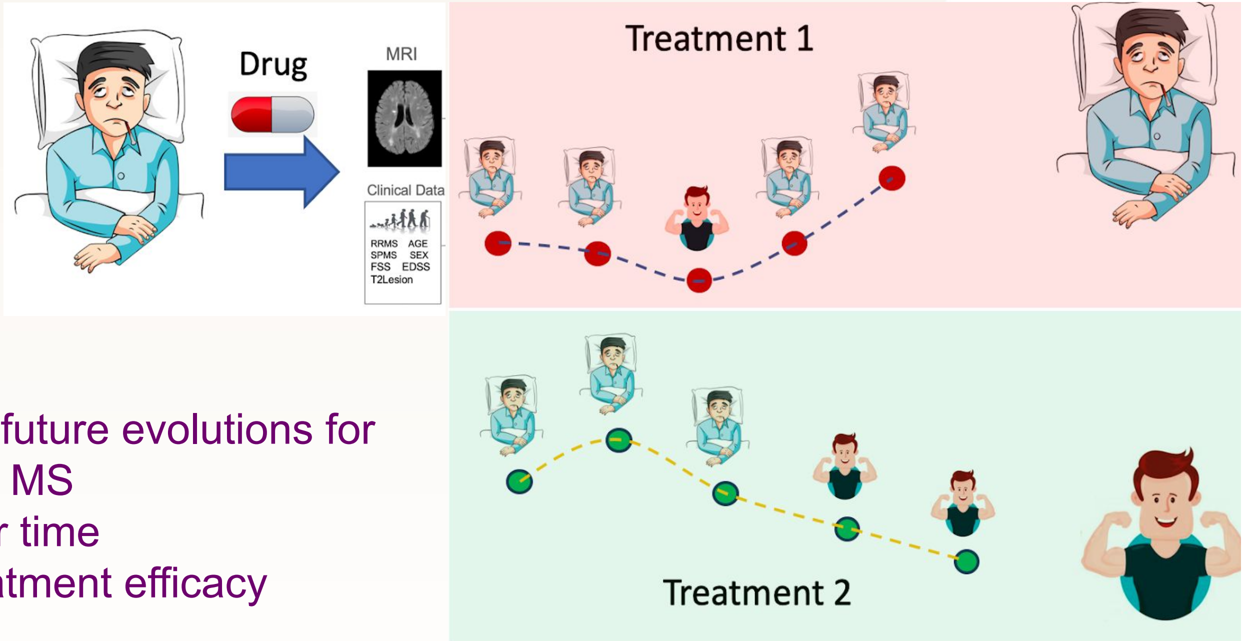


Durso-Finley *et. al*, "Improving Image-Based Precision Medicine with Uncertainty-Aware Causal Models", MICCAI 2023 **[Shortlist Best Paper]**

McGill UNIVERSITY · CIM CENTRE FOR INTELLIGENT MACHINES · PVG · Mila

Temporal Prediction of Continuous Disease Trajectories and Treatment Effects

→ First AI model to learn disease evolution over time based on medical images



- Accurate predictions of future evolutions for disability progression in MS
- Models uncertainty over time
- Increased trust and treatment efficacy

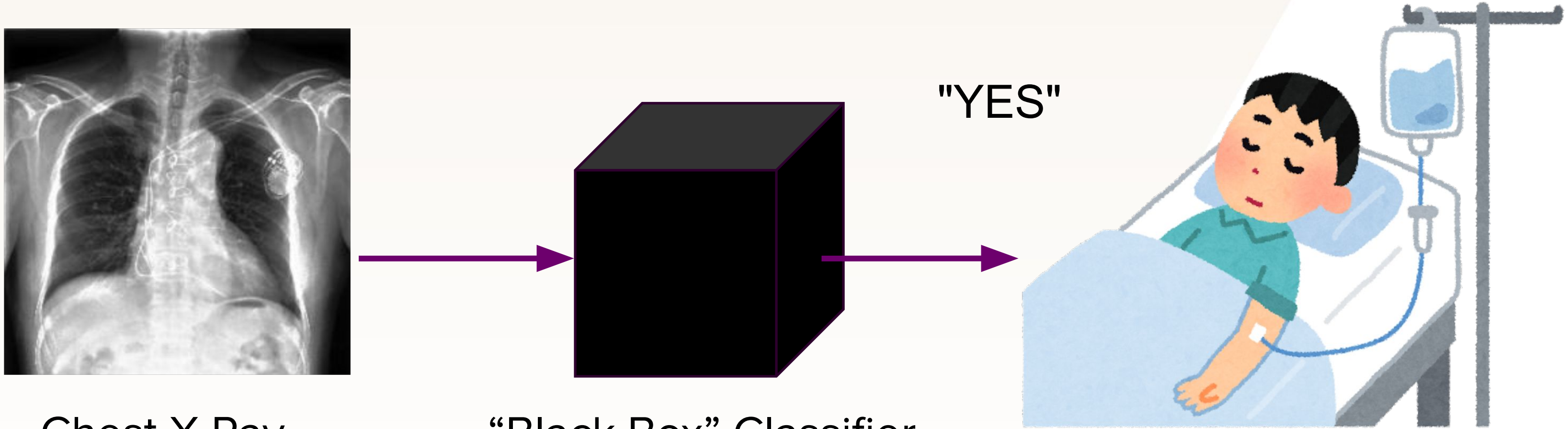
Durso-Finley, Barile B. *et. al*, "Probabilistic Temporal Prediction of Continuous Disease Trajectories and Treatment Effects Using Neural SDEs", MICCAI 2024

Trustworthy Image-Based Personalized Medicine

- Uncertainty estimation
- Explainability
- Improving biases

Black box nature of deep learning models

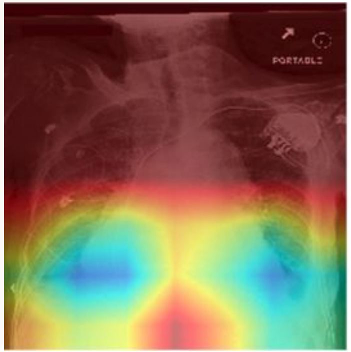
Does this patient have pleural effusion?



Chest X-Ray

“Black Box” Classifier

Opening up the black box



Where was the model looking at when it made its prediction?

→ ***What** are the specific image markers that are indicative of pleural effusion?*

Hang *et. al*, Deep transfer learning to quantify pleural effusion severity in chest X-rays, BMC 2022

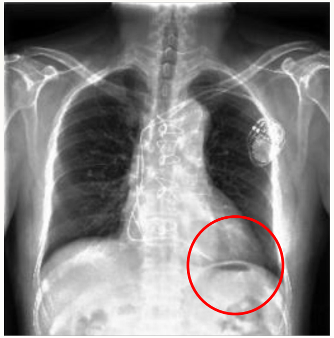
Explainability via Counterfactual Synthesis

→ This person is sick. What would the person look like if they were healthy?

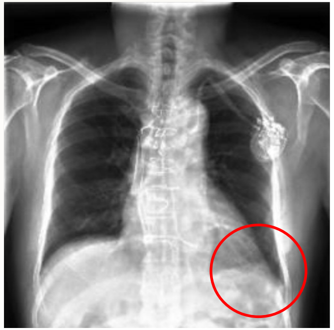
→ Goal: Counterfactual images :

- Show minimal changes
- Belongs to the correct target class
- Show realistic changes

Factual (Original) image: Pleural Effusion



Counterfactual (Synthesized) image: Healthy



Power of Vision-Language Models (VLMs)

→ Vision-Language Foundation Models have been very successful at synthesizing high-resolution images

picture of an astronaut on mars with a rover

horse running by the side of the sea

PRISM: Fine-Tuning Vision-Language Foundation Models

Pretrained Image Encoder
(Trained on diverse internet images)

Pretrained CLIP Encoder
(Trained on enormous text corpus)

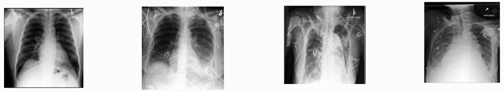
"A painting of flowers in vase "

"A photo of a sunset"

"A close-up image of a white cat"

"Eagle soaring in the sky"

Fine Tune on Medical Images
(Chest X-ray Images)

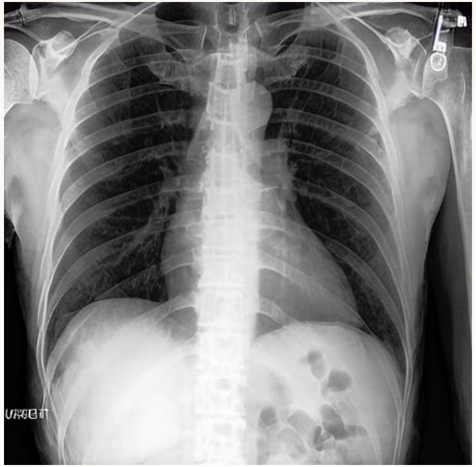


Kumar et. al, PRISM: High-Resolution & Precise Counterfactual Medical Image Generation, MIDL 2025

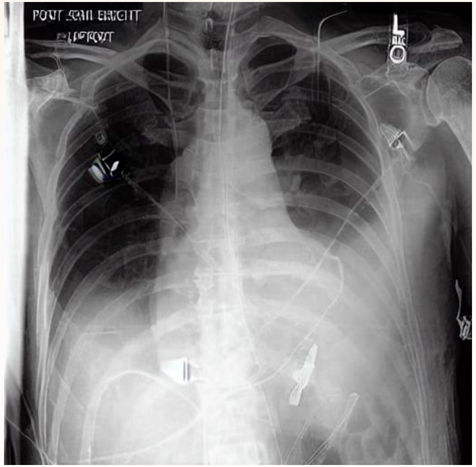
PRISM: Fine-Tuning Vision-Language Foundation Models

Chest x-ray of a patient with...

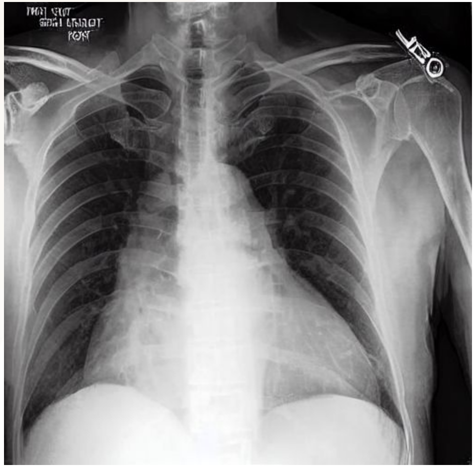
No known disease



pleural effusion and
support devices



cardiomegaly and
no devices



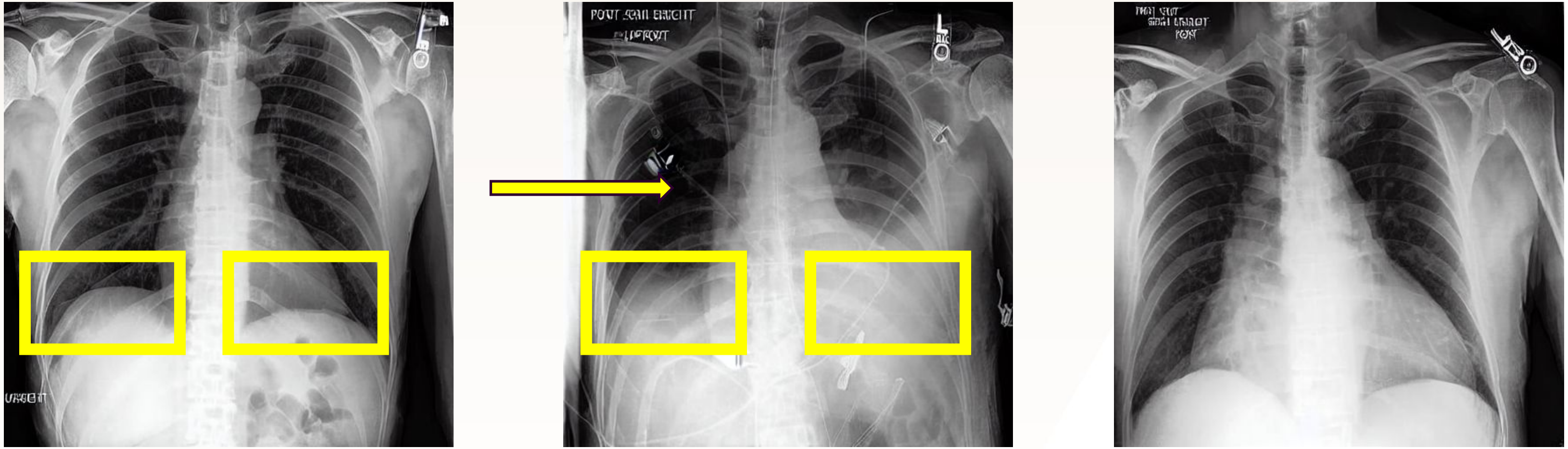
Samples are generated via finetuned Stable Diffusion v1.5

Kumar et. al, PRISM: High-Resolution & Precise Counterfactual Medical Image Generation, MIDL 2025

PRISM: Fine-Tuning Vision-Language Foundation Models

Chest x-ray of a patient with...

No known disease | pleural effusion and support devices | cardiomegaly and no devices



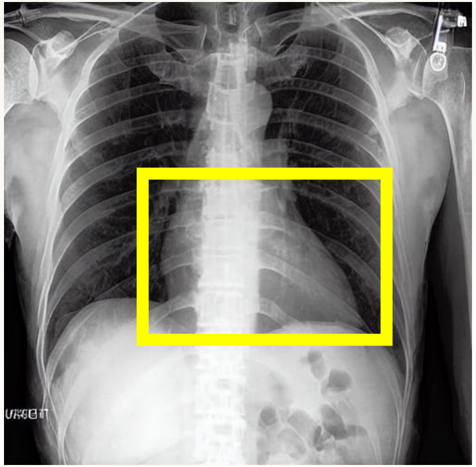
Samples are generated via finetuned Stable Diffusion v1.5

Kumar et. al, PRISM: High-Resolution & Precise Counterfactual Medical Image Generation, MIDL 2025

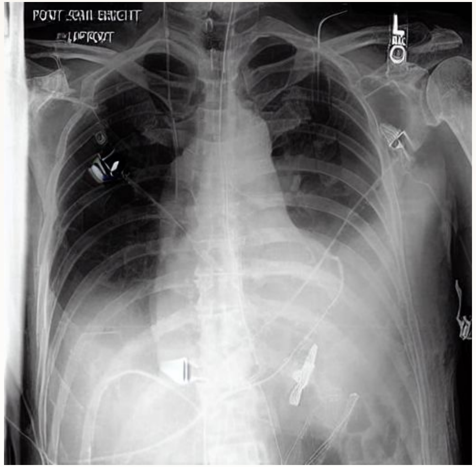
PRISM: Fine-Tuning Vision-Language Foundation Models

Chest x-ray of a patient with...

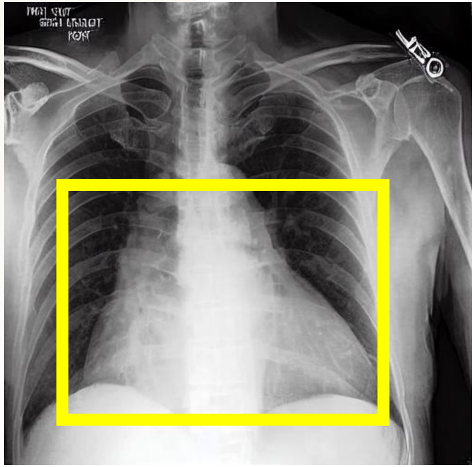
No known disease



pleural effusion and support devices



cardiomegaly and no devices



Samples are generated via finetuned Stable Diffusion v1.5

Can we leverage PRISM for CF medical image generation?

Kumar et. al, PRISM: High-Resolution & Precise Counterfactual Medical Image Generation, MIDL 2025

PRISM: The Power of VLMs for CF Medical Image Generation

PRISM: *First* method to:

- Synthesize counterfactual image at high-resolution *(512 x 512)*
- Make *precise* changes to the original image guided by natural language

```
Chest x-ray of a patient with...
```

No Disease
Sex: Male

No Disease => Pleural Effusion

Male => Female

Let's see how it does...

 Kumar, Kriz *et. al*, "Leveraging Vision-Language Foundation Models to Reveal Hidden Image-Attribute Relationships in Medical Imaging", MIV-CVPR Proceedings 2025

PRISM: The Power of VLMs for CF Medical Image Generation

Factual (Original) image

No Disease
Sex: Male

CF (Synthesized) images

No Disease =>
Pleural Effusion

Male => Female

SOTA [Non VLM]: VAE | **PRISM [Our Method]**

Pleural Effusion; Sex: Male

No Disease; Sex: Female

Kumar, Kriz *et. al*, "Leveraging Vision-Language Foundation Models to Reveal Hidden Image-Attribute Relationships in Medical Imaging", MIV-CVPR Proceedings 2025

PRISM: The Power of VLMs for CF Medical Image Generation

CF (Synthesized) images

SOTA [Non VLM]: VAE | PRISM [Our Method]

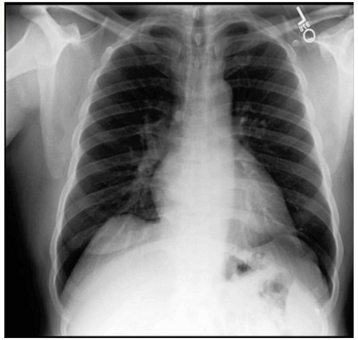
Factual (Original) image

No Disease
Sex: Male

No Disease =>
Pleural Effusion

Pleural Effusion; Sex: Male

Male => Female

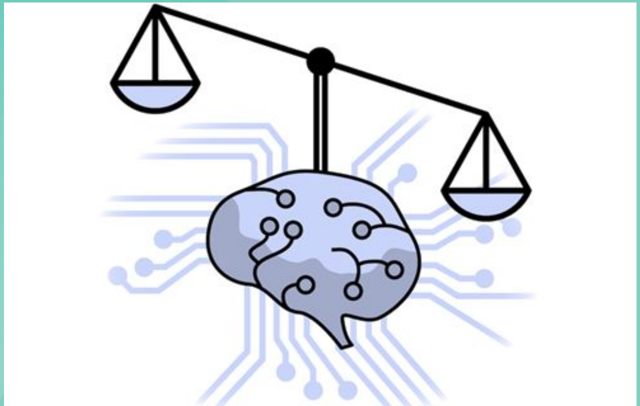


No Disease; Sex: Female

 Kumar, Kriz *et. al*, “Leveraging Vision-Language Foundation Models to Reveal Hidden Image-Attribute Relationships in Medical Imaging”, MIV-CVPR Proceedings 2025

Trustworthy Image-Based Personalized Medicine

- Uncertainty estimation
- Explainability
- Improving biases



Medical Imaging Biases – Beyond Known Attributes

Training samples

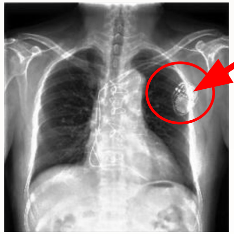
Example:

Patients with Pleural Effusion typically have variety of medical devices:

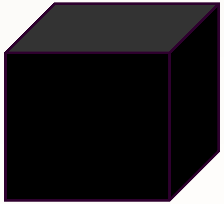
- Chest drains; wires; tubes; pacemakers

Inference

pacemaker



Patient: No Pleural Effusion



Disease Classifier

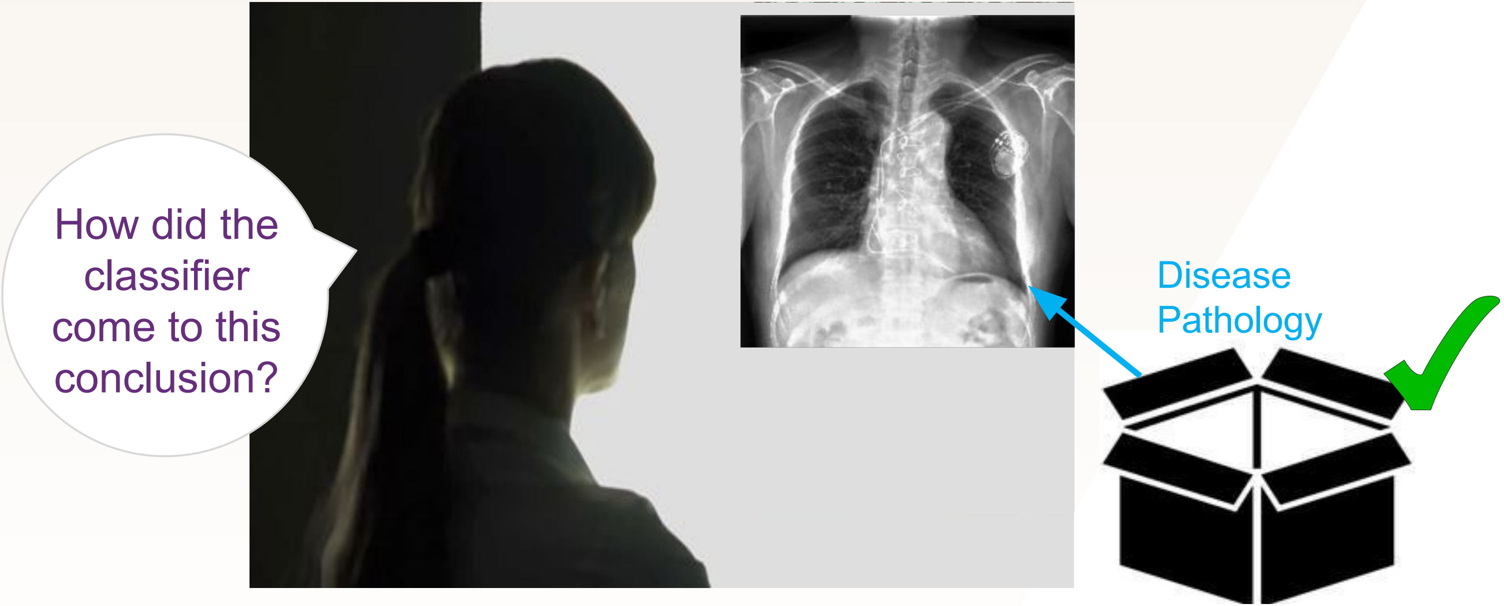
Wrong!

Pleural Effusion

Latched onto visual artifact: pacemaker!

Exposing Medical Image Biases

Classifier: Pleural Effusion



Mila

Exposing Medical Image Biases

Classifier: Pleural Effusion

How did the classifier come to this conclusion?

Wrong reason!

Visual Artifact (Shortcut)

→ Question: Can we get rid of all medical devices?

Mila

A Difficult Task: Removing Medical Devices

→ SOTA Counterfactual Images synthesized prior to PRISM

DeCoDex (SOTA before PRISM) struggled with removing all devices

Factual (Original) image
Patients with Medical Devices

Counterfactual (Synthesized) image
Patients without Medical Devices

Medical Devices:
Chest drains; tubes; pacemakers

→ Can PRISM remove all support devices using a single language prompt?

Fathi, Kumar *et. al*, “DeCoDEx”, MIDL 2024 **[Shortlist Best Paper]**

Mila

PRISM: Removing Support Devices with Ease

Chest x-ray of a patient with a lot medical devices

→ Chest x-ray of a patient without medical devices

→ Easily removes all medical devices!
No specific training labels for wires, pacemakers, etc.

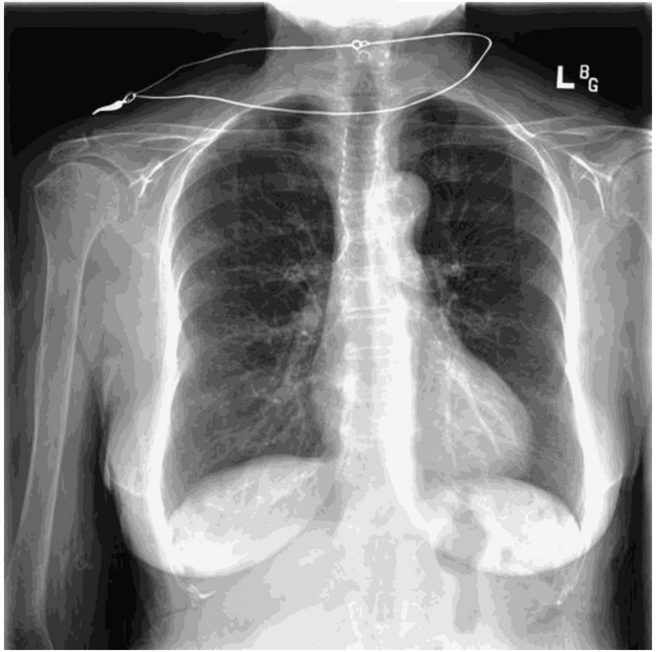
Kumar et. al, PRISM: High-Resolution & Precise Counterfactual Medical Image Generation, MIDL 2025

PRISM: Explainability & Debiasing with VLMs

→ Does PRISM ignore all support devices when generating a CF image of a patient with a disease?

`Chest x-ray of a patient with cardiomegaly →`

`Chest x-ray of a patient with no finding`



→ Focuses on relevant features and ignores the confounding artifacts (wires)

Kumar et. al, PRISM: High-Resolution & Precise Counterfactual Medical Image Generation, MIDL 2025

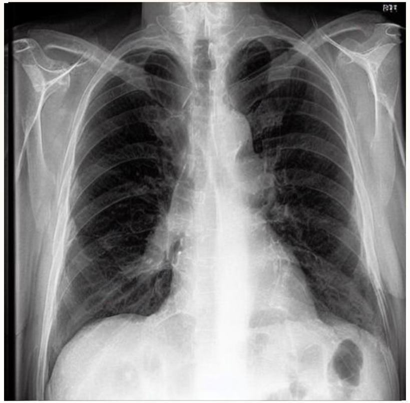
Mila

PRISM: Discovering Hidden Relations Between Attributes

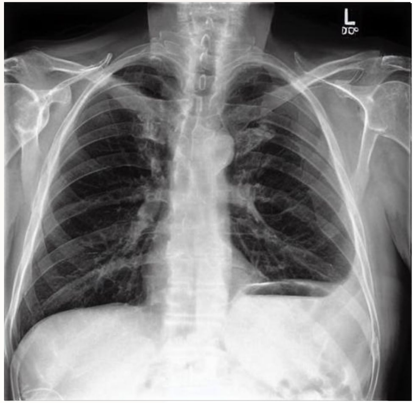
→ *Does PRISM discover relationships between attributes, without explicitly training on them?*

`Chest x-ray of a patient with cardiomegaly` →
`Chest x-ray of a patient with no finding`

Patient 1



Patient 2



- *Removes disease & pacemaker*
- *Wasn't told about pacemaker!*

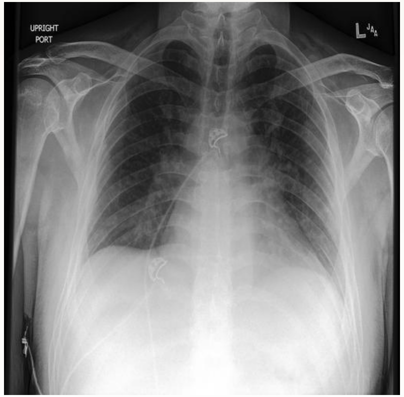
➡ *Turns out pacemaker is a treatment for cardiomegaly!*

 Kumar, Kriz *et. al*, "Leveraging Vision-Language Foundation Models to Reveal Hidden Image-Attribute Relationships in Medical Imaging", MIV-CVPR Proceedings 2025

Technology Revolution – Fast Moving Age of Wonder

Achieving state of the art on new tasks *daily*

🩺Image generation
🩺Report (diagnosis) generation
🩺Visual Question Answering



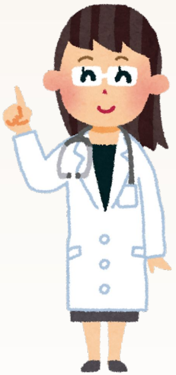
Question: Does this image have pleural effusion?

Answer: Yes, this chest X-ray shows evidence of pleural effusion. There appears to be blunting of the right costophrenic angle and increased opacity in the lower right hemithorax, which are characteristic radiographic signs of pleural effusion. The effusion appears to be unilateral (on the right side) and moderate in size based on this frontal (PA/AP) view.

Mila

Promise of Modern AI for Clinical Decision Support

Provide clinicians with <u>trustworthy</u> AI tools to predict future individual treatment response on different treatments using *multimodal data*



Genomics EHR Lab Tests

Multimodal Patient Data

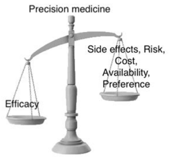
Secure 🔒 Validated √

Unbiased ⚖️ Explainable 💡

Trustworthy AI system

Treatment A: 40% Treatment B: 60%

Treatment Response Prediction



Causality & Reasoning

Generative Models for 3D images

Temporal Disease Models

Mila

Thank you for your attention!

Probabilistic Vision Group

Collaborators

Douglas L. Arnold, Montreal Neurological Institute
Nick Powlowski, MSR
Mohammed Havaei, Google Research

Sponsors

Google Research
Multiple Sclerosis Society of Canada
Mila
ILLS International Laboratory on Learning Systems
NSERC CRSNG
CIFAR CANADIAN INSTITUTE FOR ADVANCED RESEARCH
ICRA INSTITUT CANADIEN DE RECHERCHES AVANCÉES
Microsoft Research
INTERNATIONAL PROGRESSIVE MS ALLIANCE
CONNECT TO END PROGRESSIVE MS

This work was made possible by Biogen, BioMS, MedDay, Novartis, Roche / Genentech, and Teva who generously provided the data

Probabilistic Vision Group

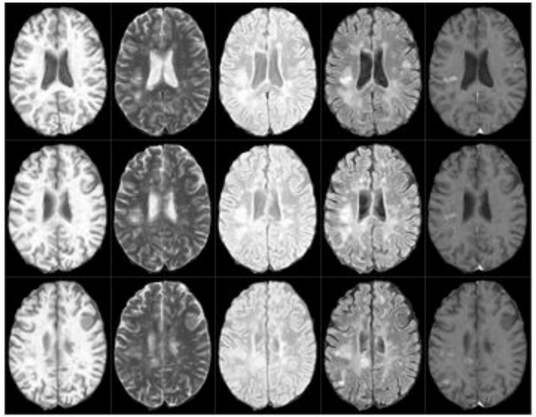
Extra Slides

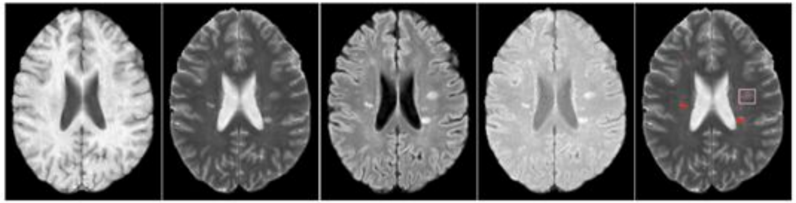
"Big Data" : MS Clinical Trials

- First federated clinical and MRI dataset - large phase 3 MS clinical trials

- Tens of thousands of patients scans
  - At multiple global imaging centres
  - On multiple scanner technologies

- Large MRI volumes
  - Multiple imaging modalities
  - At multiple timepoints
  - With expert-generated lesion labels





Patient on different treatments of varying efficacy and on placebos

McGill UNIVERSITY
CIM CENTRE FOR INTELLIGENT MACHINES
PVG
Mila

Earlier Methods – VAEs, CycleGANs

CF (Synthesized) image (SOTA)

Factual (Original) image

Finding: No Disease
Sex: Male

Change Finding= Pleural Effusion

Finding: Pleural Effusion
Sex: Male

Change sex = Female

Finding: No Disease
Sex: Female

Ribeiro *et. al*, "High Fidelity Image Counterfactuals with Probabilistic Causal Models", ICML 2023

Mila

Earlier Methods – VAEs, CycleGANs

CF (Synthesized) image (SOTA)

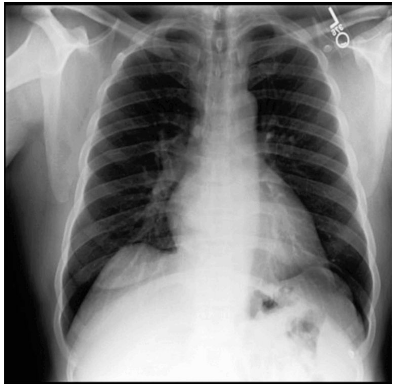
Factual (Original) image

Finding: No Disease
Sex: Male

Change Finding= Pleural Effusion

Change sex = Female

Finding: Pleural Effusion
Sex: Male



Finding: No Disease
Sex: Female

Ribeiro *et. al*, “High Fidelity Image Counterfactuals with Probabilistic Causal Models”, ICML 2023

PRISM: Removing Support Devices with Ease

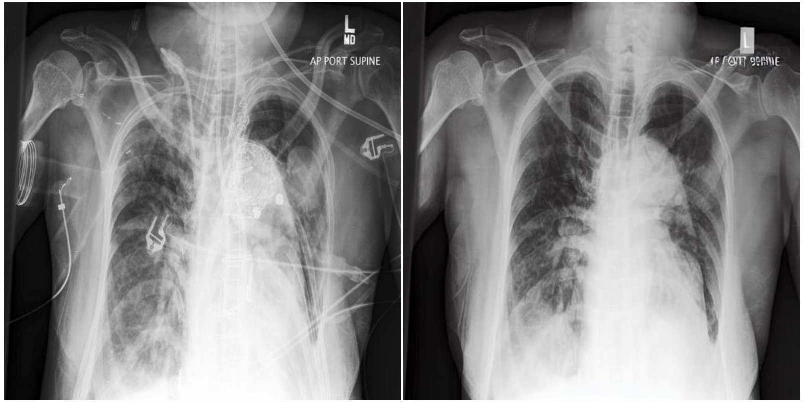
More examples....

`Chest x-ray of a patient with a lot of medical devices`

`→ Chest x-ray of a patient without medical devices`

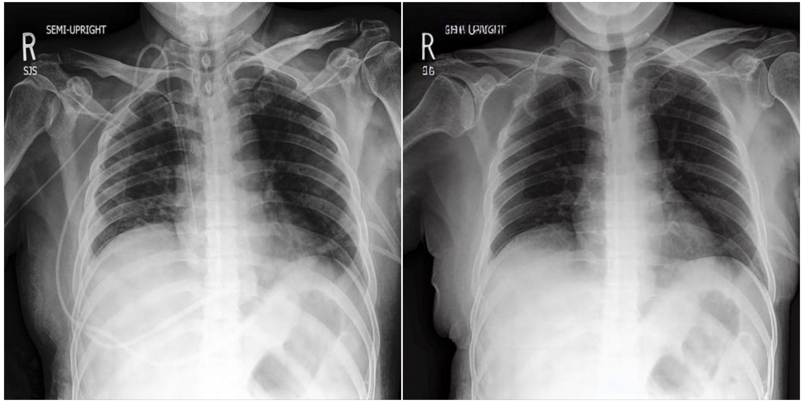
All types of devices removed with a single text prompt!

Before: Devices | After: No Devices



Patient 1

Before: Devices | After: No Devices



Patient 2

Kumar et. al, PRISM: High-Resolution & Precise Counterfactual Medical Image Generation, MIDL 2025

VLMs for Medical Image Generation

Did the foundation model already learn all about radiological images?

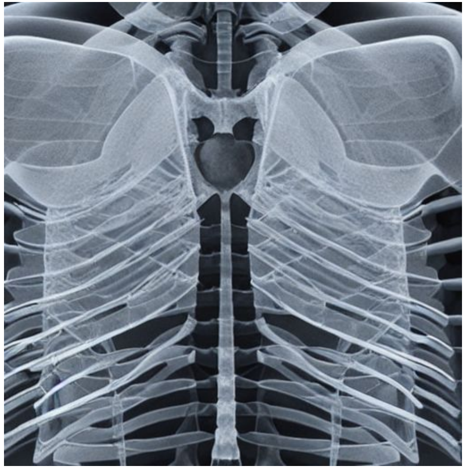
What if we didn't fine-tune to our medical imaging dataset?

VLMs for Medical Image Generation

Did the foundation model already learn all about radiological images?

What if we didn't fine-tune to our medical imaging dataset?

"Chest x-ray of a patient with medical devices"



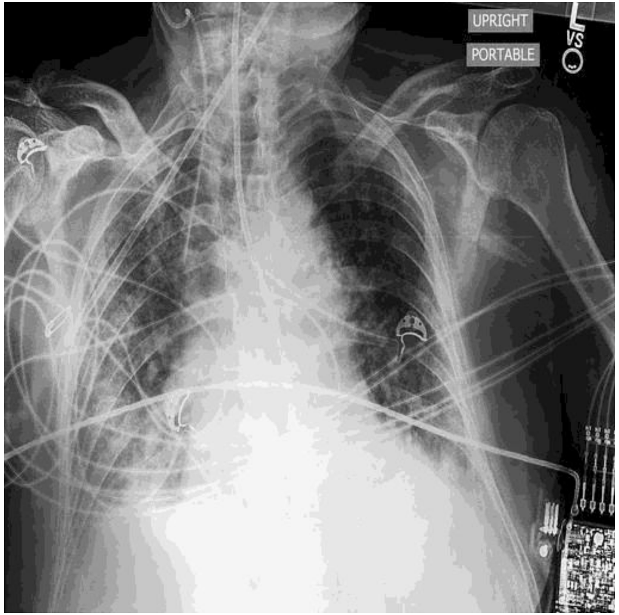
They don't seem to know much about radiological images on their own...

PRISM: Explainability & Debiasing with VLMs

→ Can VLM ignore all support devices when generating a CF image?

```
Chest x-ray of a patient
with pleural effusion →

Chest x-ray of a patient
with no finding
```



Kumar et. al, PRISM: High-Resolution & Precise Counterfactual Medical Image Generation, MIDL 2025