Lessons from Deploying AI in Healthcare

James Zou

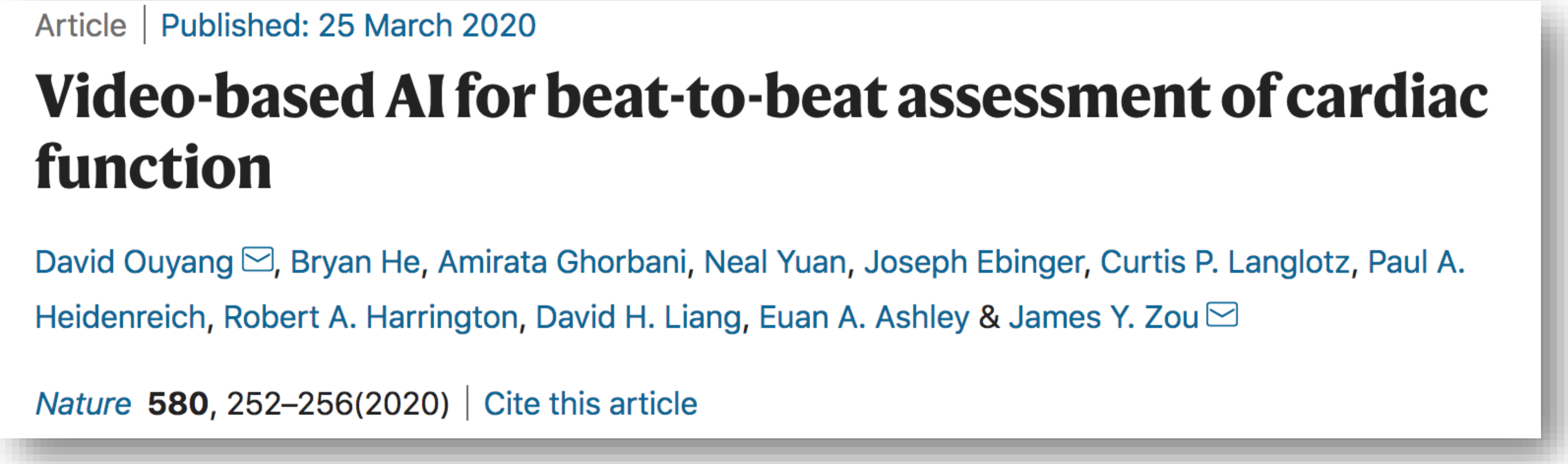
Stanford University

March 5, 2022

www.james-zou.com

jamesz@stanford.edu

Example 1: deploying cardiology AI

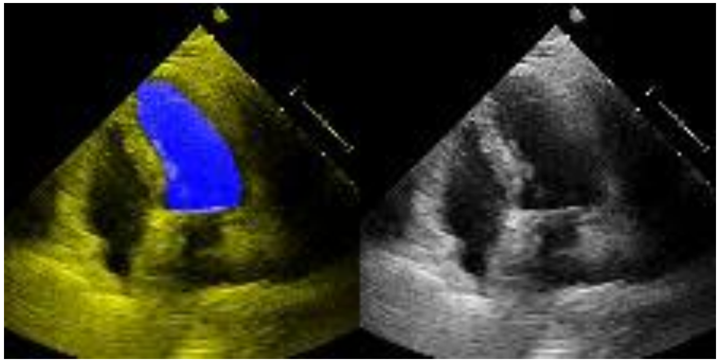


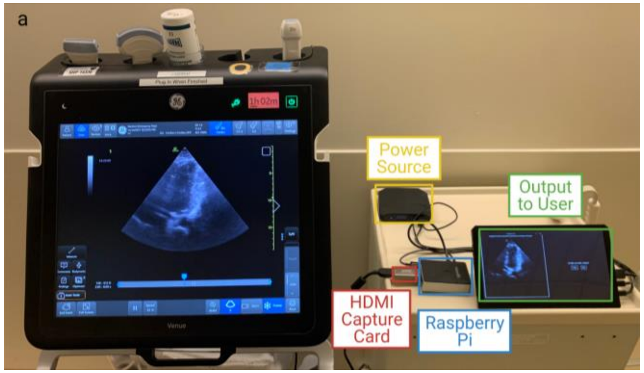
Article | Published: 25 March 2020

Video-based AI for beat-to-beat assessment of cardiac function

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley & James Y. Zou

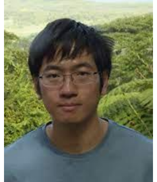
Nature **580**, 252–256(2020) | Cite this article





David Ouyang

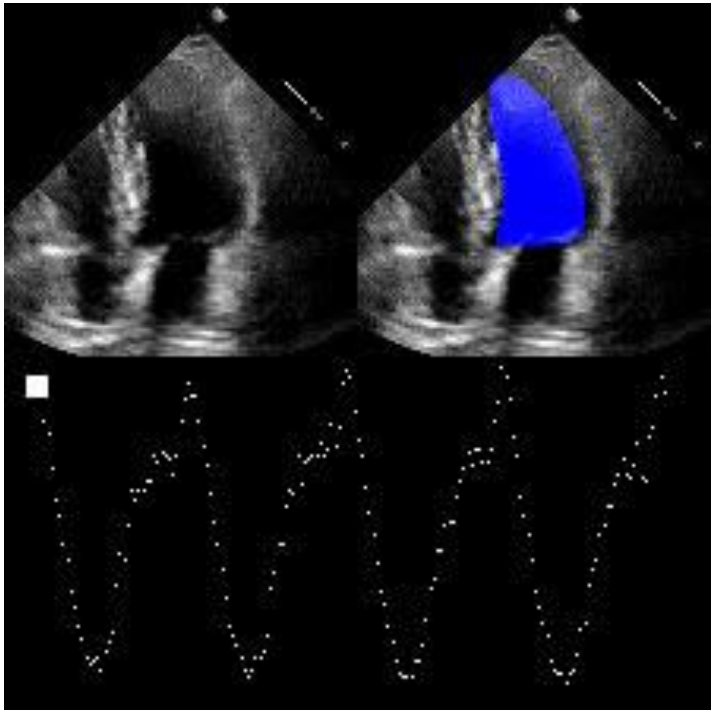
Bryan He



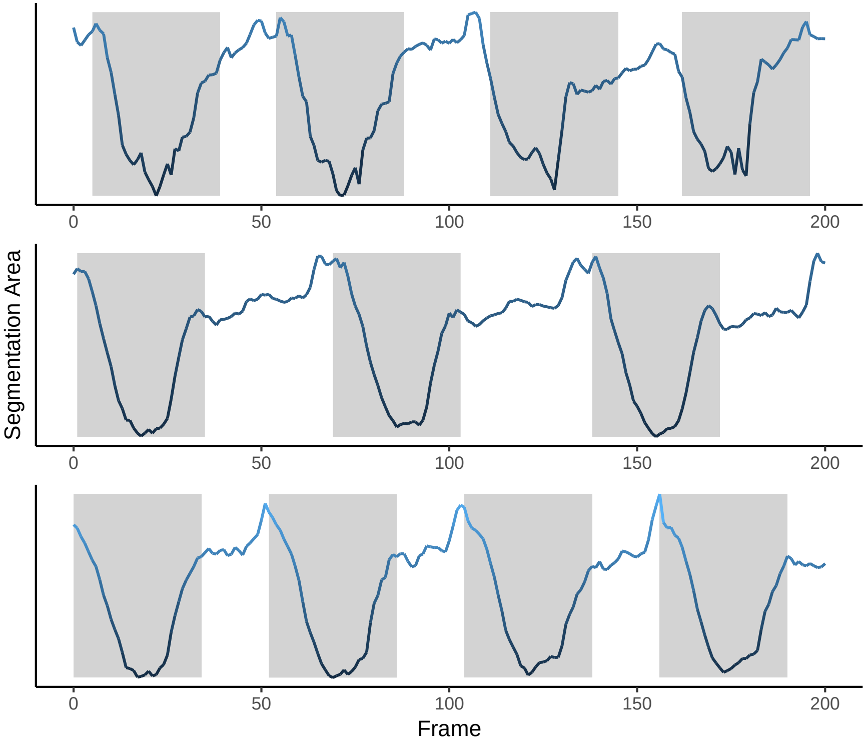
Ouyang et al. *Nature* 2020

Computer vision assesses cardiac ultrasound

Algorithm output



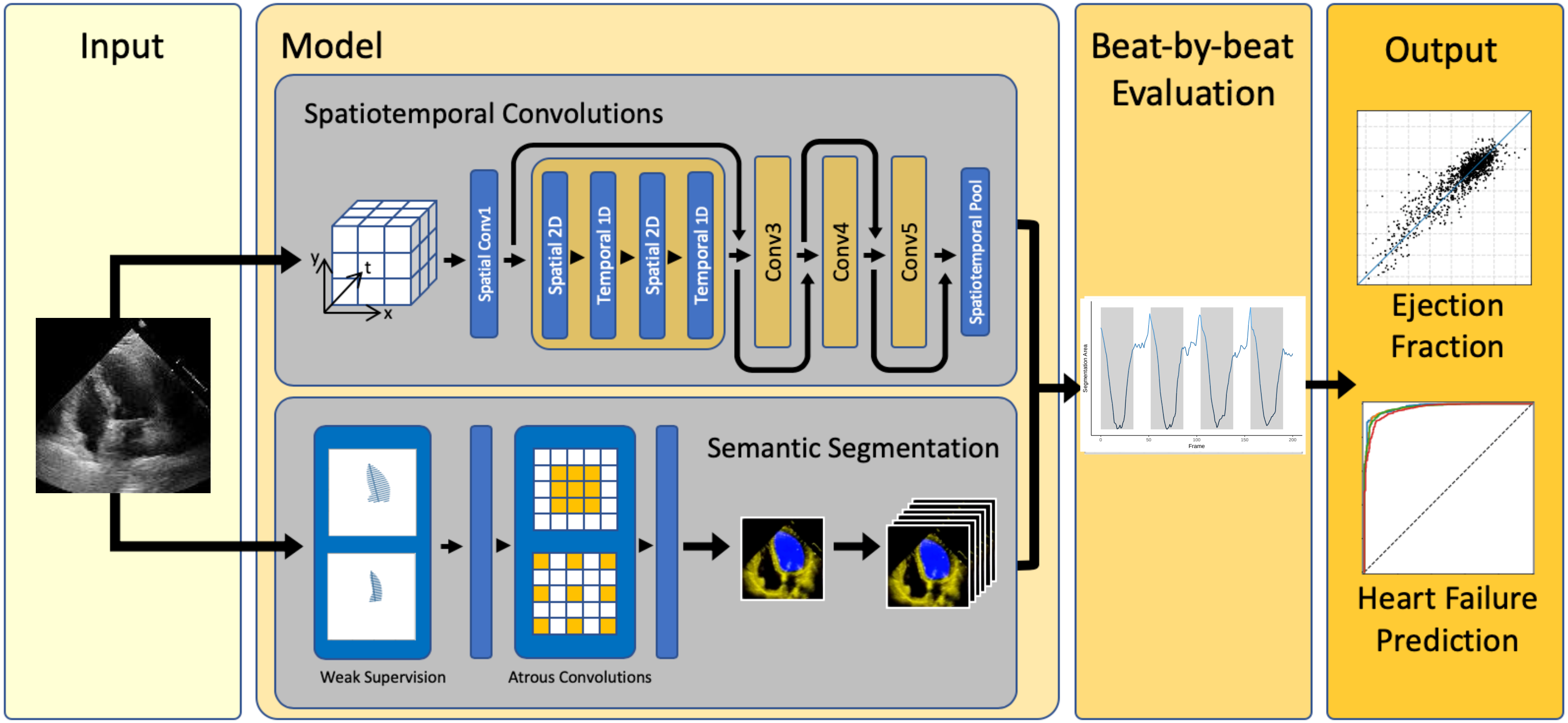
EchoNet assessed chamber area



Ouyang et al. *Nature* 2020

Algorithm mimics clinical workflow

EchoNet-Dynamic



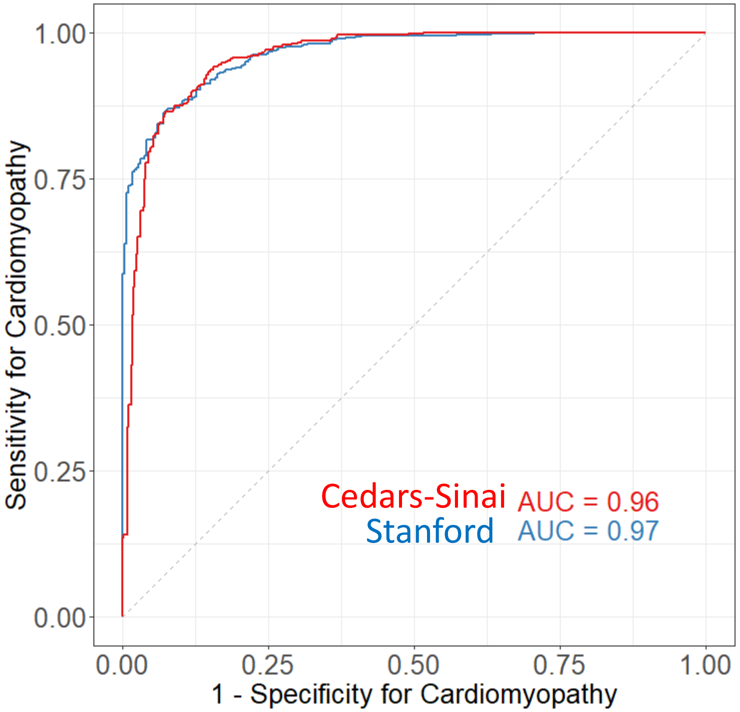
Kidney
Liver
…

Idea: use temporal segmentation to focus attention of model.

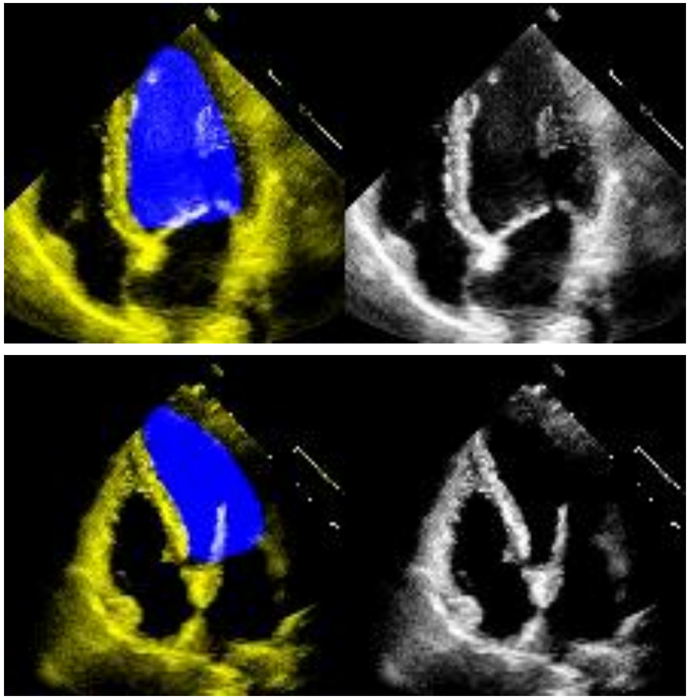
Ouyang et al. *Nature* 2020

Achieves expert performance in new hospital

Predicting heart failure



Examples



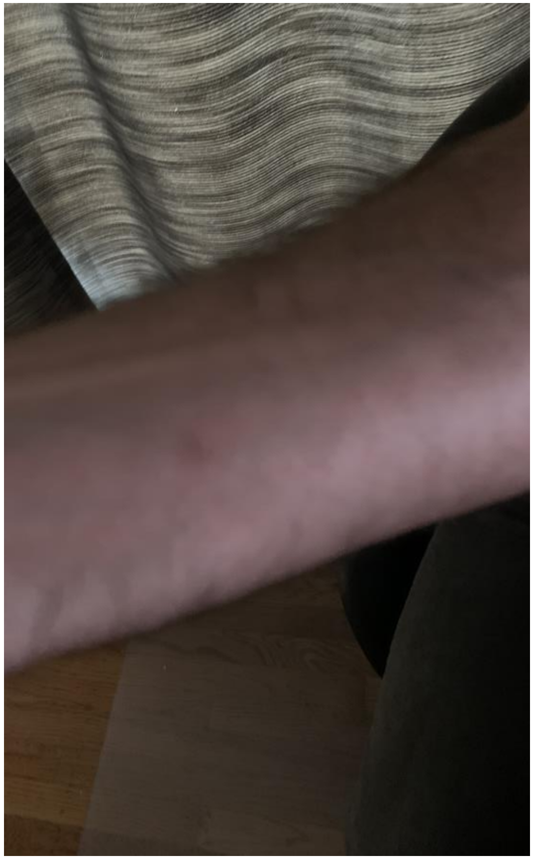
Ouyang et al. *Nature* 2020

Example 2: AI to improve telemedicine

COVID-19 → **50x increase in digital visits**

Many patient photos for telemedicine are poor quality

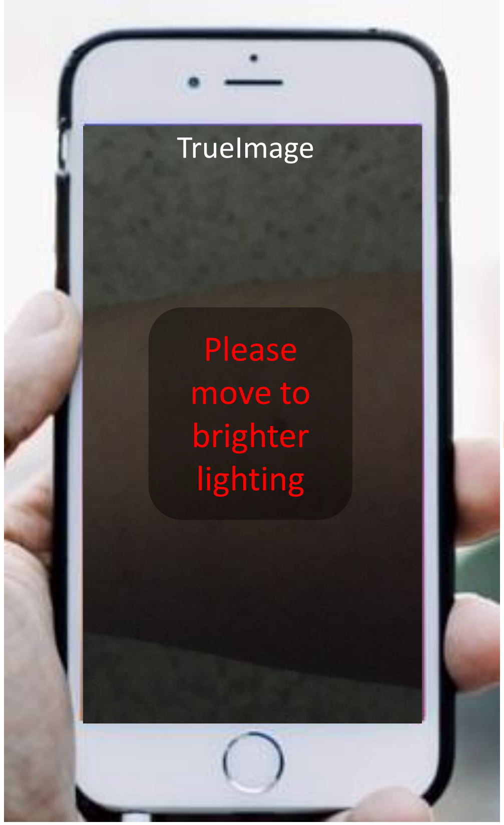
- Manual review of photos prior to the physician encounter consumed >2000 hours in 2021 at Stanford
- Poor quality photos disrupt clinic workflow
- Improving teledermatology = improving access to care



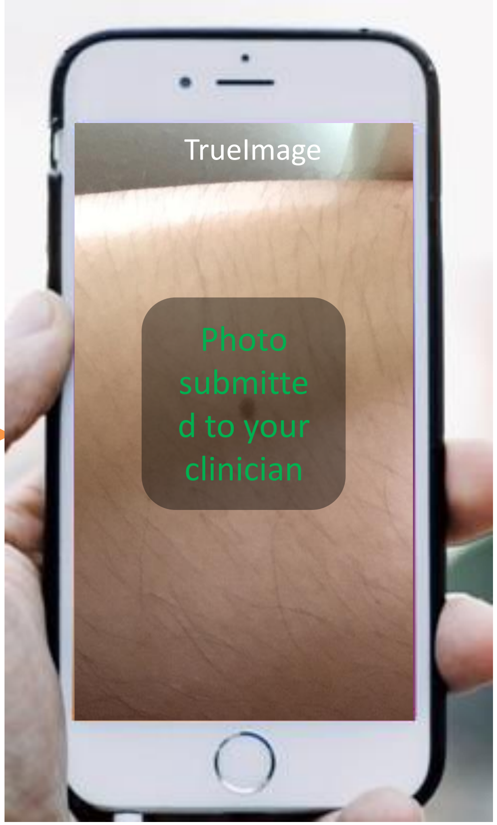
TrueImage = Online check deposit for dermatology

Vodrahalli et al, PSB 2021

TrueImage Workflow

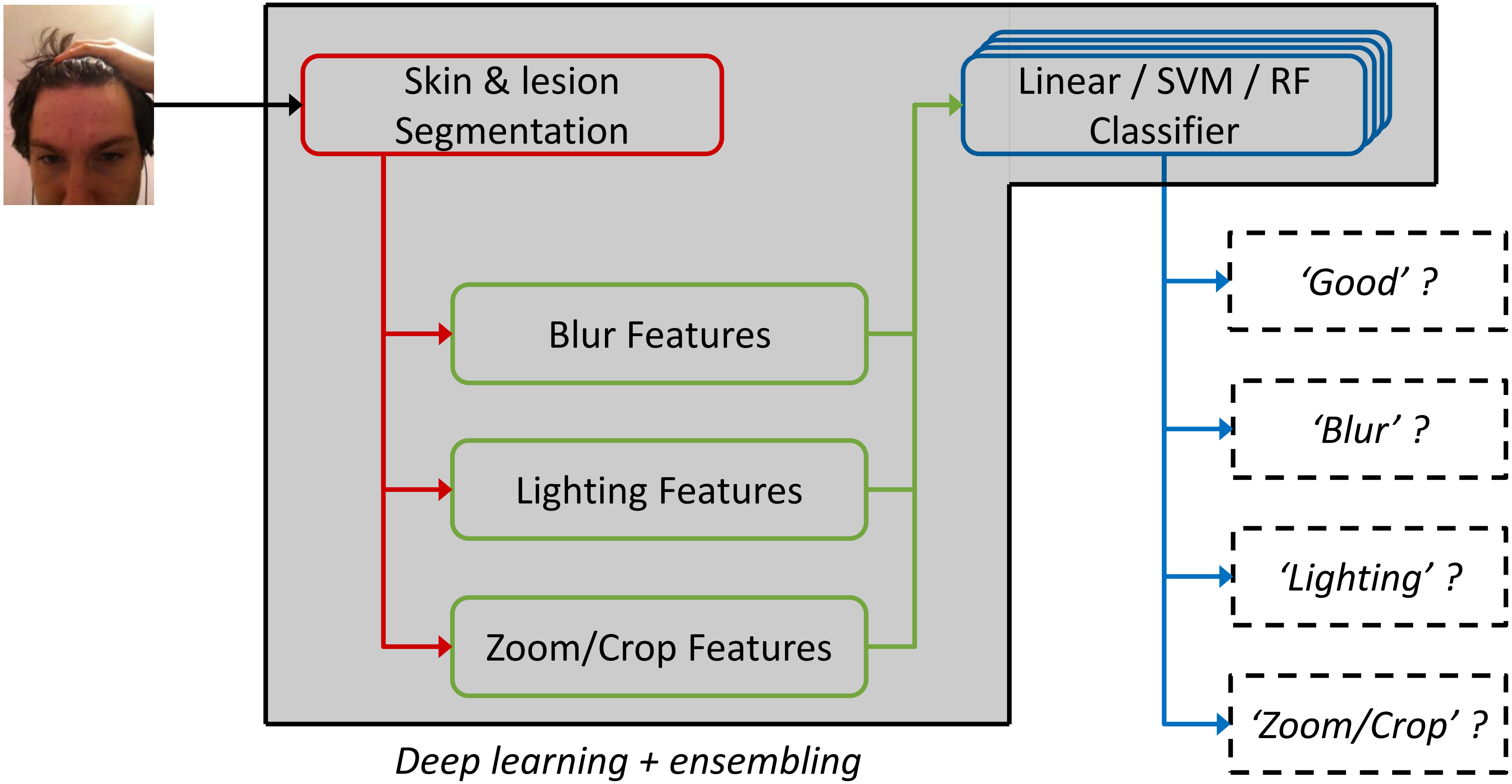


Move to brighter lighting



Vodrahalli et al, PSB 2021

TrueImage Algorithm



Skin & lesion Segmentation

Blur Features

Lighting Features

Zoom/Crop Features

Linear / SVM / RF Classifier

'Good' ?

'Blur' ?

'Lighting' ?

'Zoom/Crop' ?

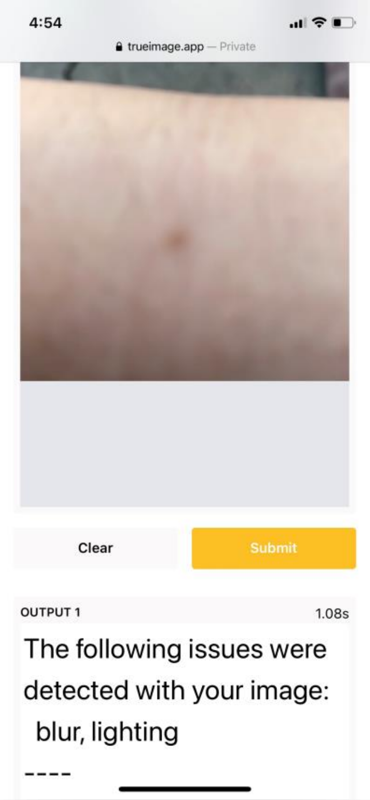
Deep learning + ensembling

Vodrahalli et al, PSB 2021

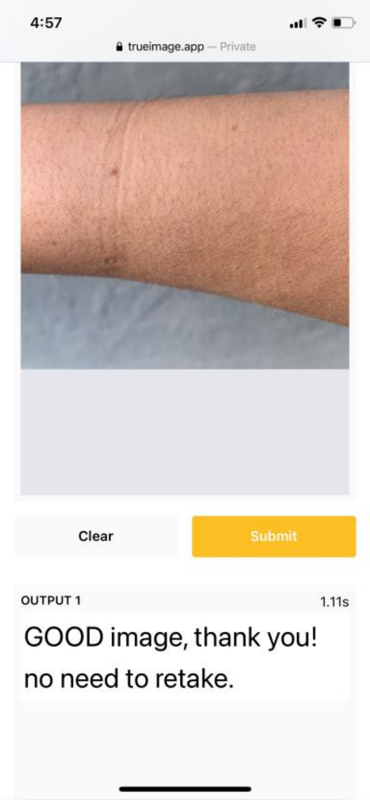
Prospective study at Stanford

TrueImage filters 80% of poor quality photos; takes <1 minute per patient

What does improvement look like?

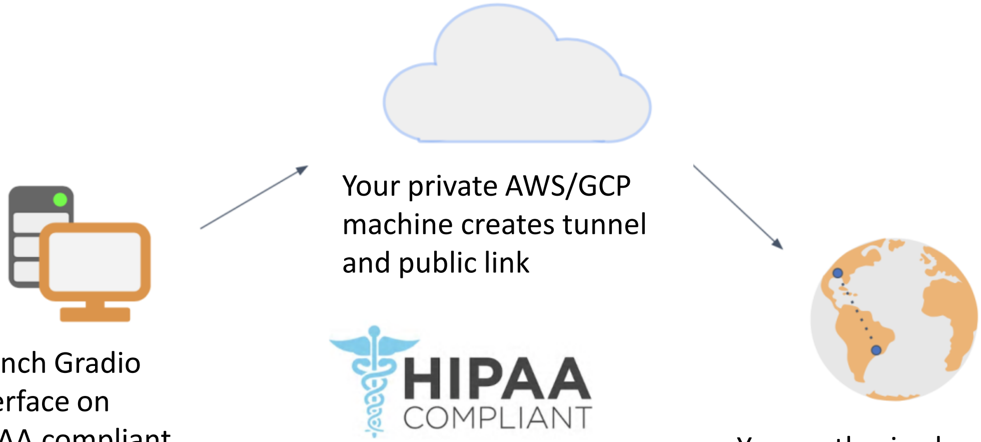


The following issues were detected with your image:
blur, lighting



GOOD image, thank you!
no need to retake.

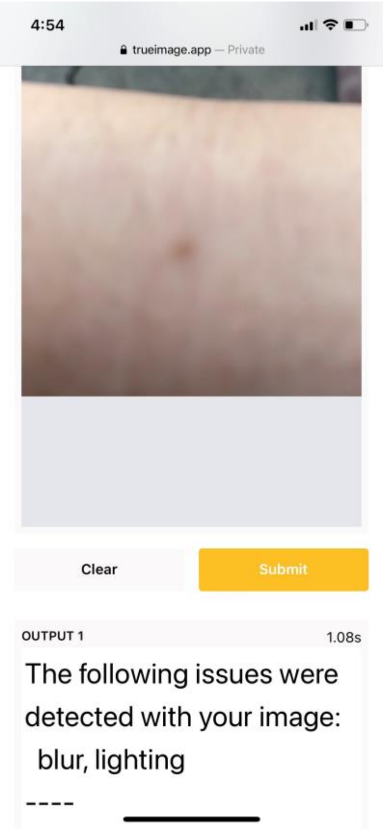
Deploying TrueImage at Stanford clinics



Launch Gradio interface on HIPAA compliant servers

Your private AWS/GCP machine creates tunnel and public link

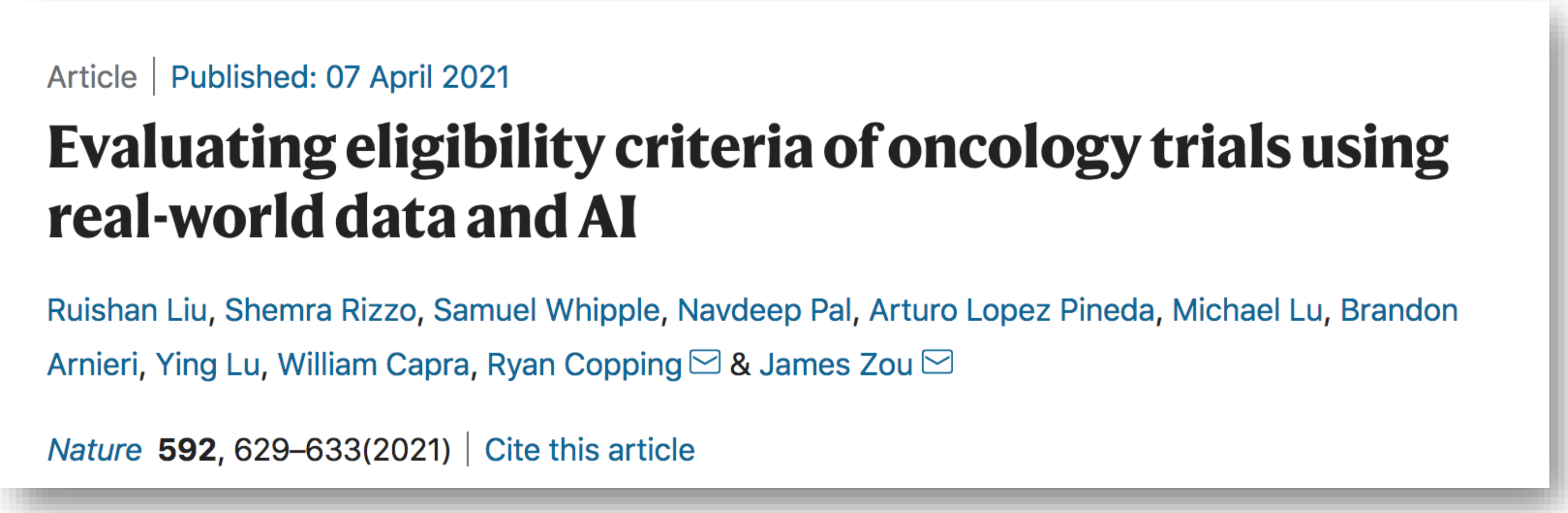
Your authorized users can now access the model



OUTPUT 1 1.08s

The following issues were detected with your image:
blur, lighting

Example 3: AI to design clinical trials

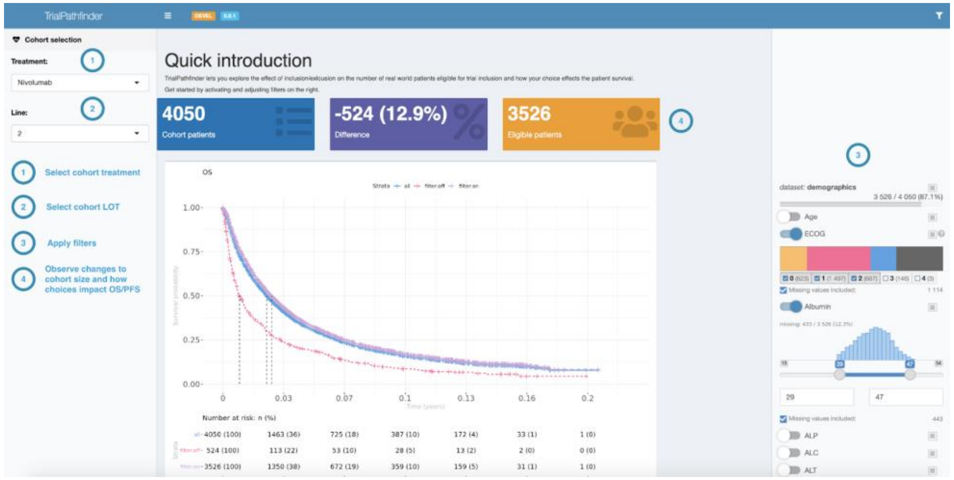


Article | Published: 07 April 2021

Evaluating eligibility criteria of oncology trials using real-world data and AI

Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping & James Zou

Nature **592**, 629–633(2021) | Cite this article



Treatment: Nivolumab

Line: 2

1 Select cohort treatment

2 Select cohort LOT

3 Apply filters

4 Observe changes to cohort size and how choices impact OS/PFS

Quick introduction

TrialPathfinder lets you explore the effect of inclusion/exclusion on the number of real world patients eligible for trial inclusion and how your choice effects the patient survival.

Get started by activating and adjusting filters on the right.

4050 Cohort patients

-524 (12.9%) Difference

3526 Eligible patients

Number at risk: n (%)

Strata							
all	4050 (100)	1463 (36)	725 (18)	387 (10)	172 (4)	33 (1)	1 (0)
filter.off	524 (100)	113 (22)	53 (10)	28 (5)	13 (2)	2 (0)	0 (0)
filter.on	3526 (100)	1350 (38)	672 (19)	359 (10)	159 (5)	31 (1)	1 (0)

dataset: **demographics**

3 526 / 4 050 (87.1%)

Age

ECOG

☑ 0 (623) ☑ 1 (1 497) ☑ 2 (667) ☐ 3 (146) ☐ 4 (3)

☑ Missing values included: 1 114

Albumin

missing: 433 / 3 526 (12.3%)

29 47

☑ Missing values included: 443

ALP

ALC

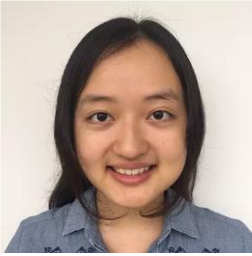
ALT

Roche

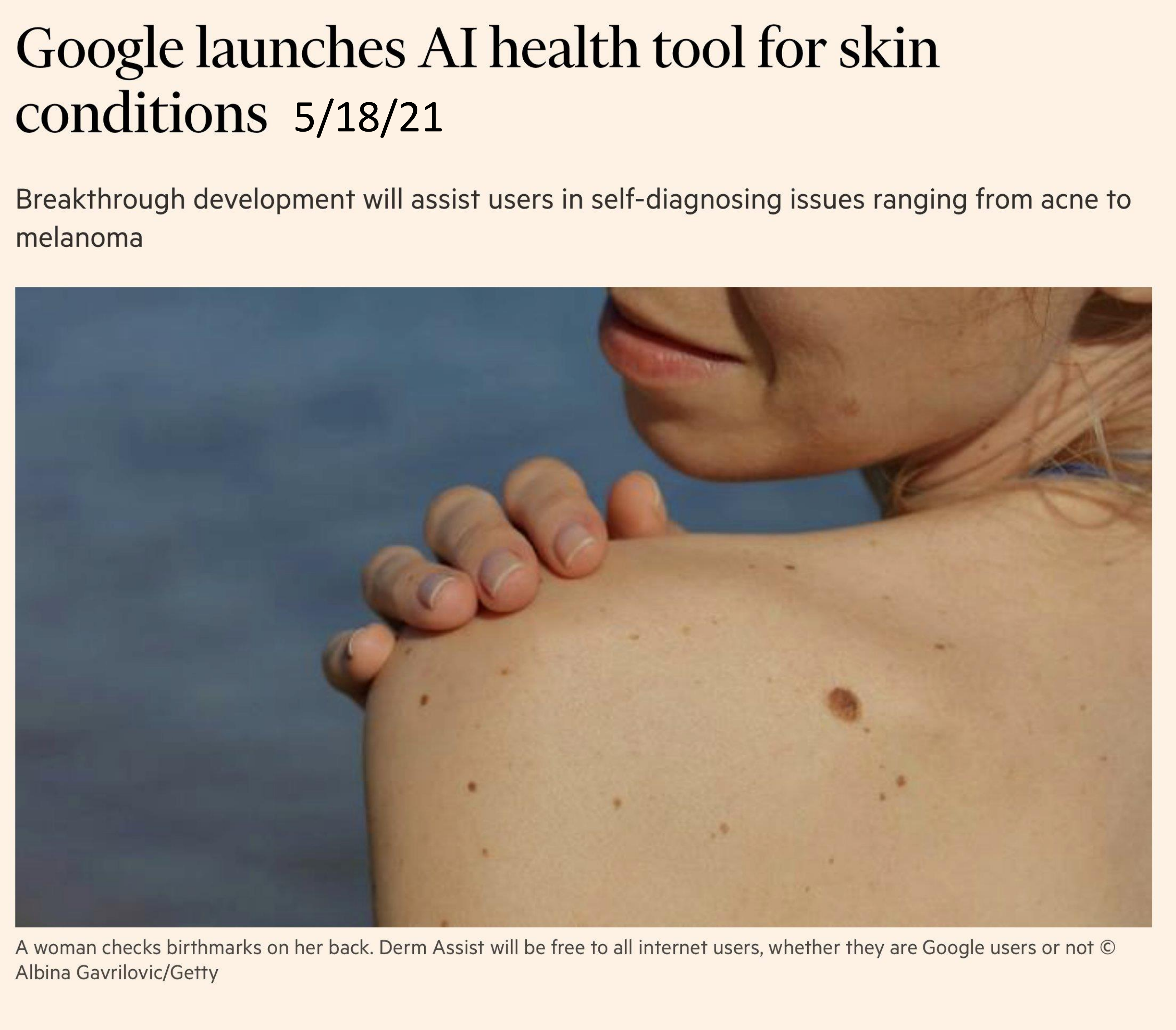
Genentech

A Member of the Roche Group

Ruishan Liu



Liu et al. *Nature* 2021

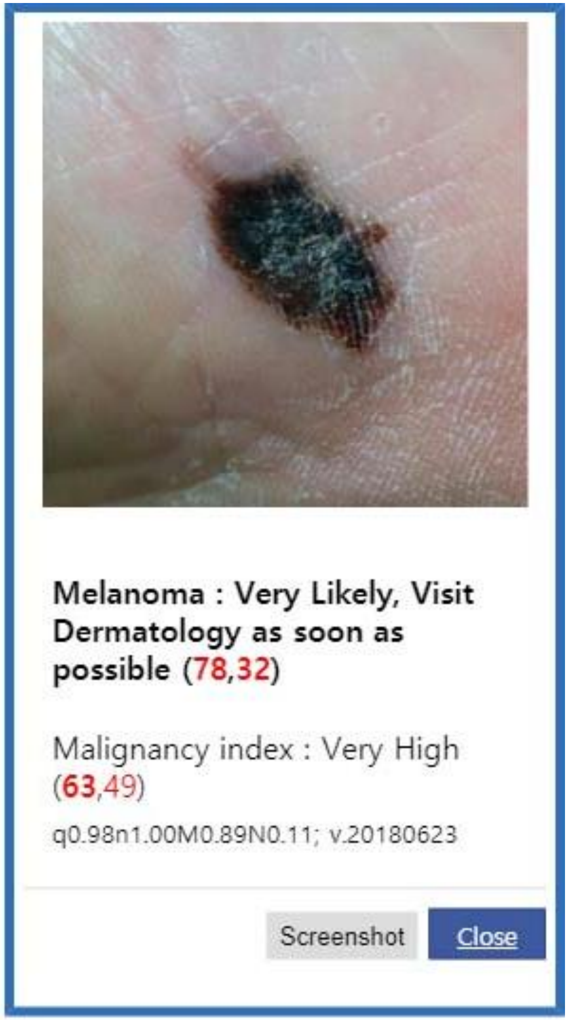


Google launches AI health tool for skin conditions 5/18/21

Breakthrough development will assist users in self-diagnosing issues ranging from acne to melanoma

A woman checks birthmarks on her back. Derm Assist will be free to all internet users, whether they are Google users or not © Albina Gavrilovic/Getty

AI dermatology apps



Melanoma : Very Likely, Visit Dermatology as soon as possible (78,32)

Malignancy index : Very High (63,49)

q0.98n1.00M0.89N0.11; v.20180623

ModelDerm

Original reported: 0.93 AUC

↓

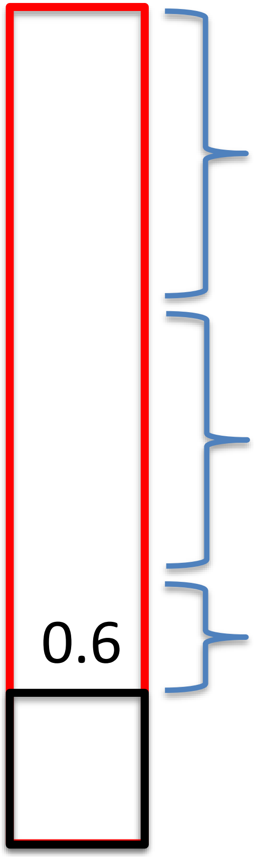
Stanford patients: 0.60 AUC

Roxana Daneshjou

Why did the Derm AI performance crater?

0.93 AUC

reported test



Stanford test

0.16 visual annot. (prev) vs. biopsy (Stanford)

label mistakes

0.12 common diseases (prev) vs. all (Stanford)

data shift

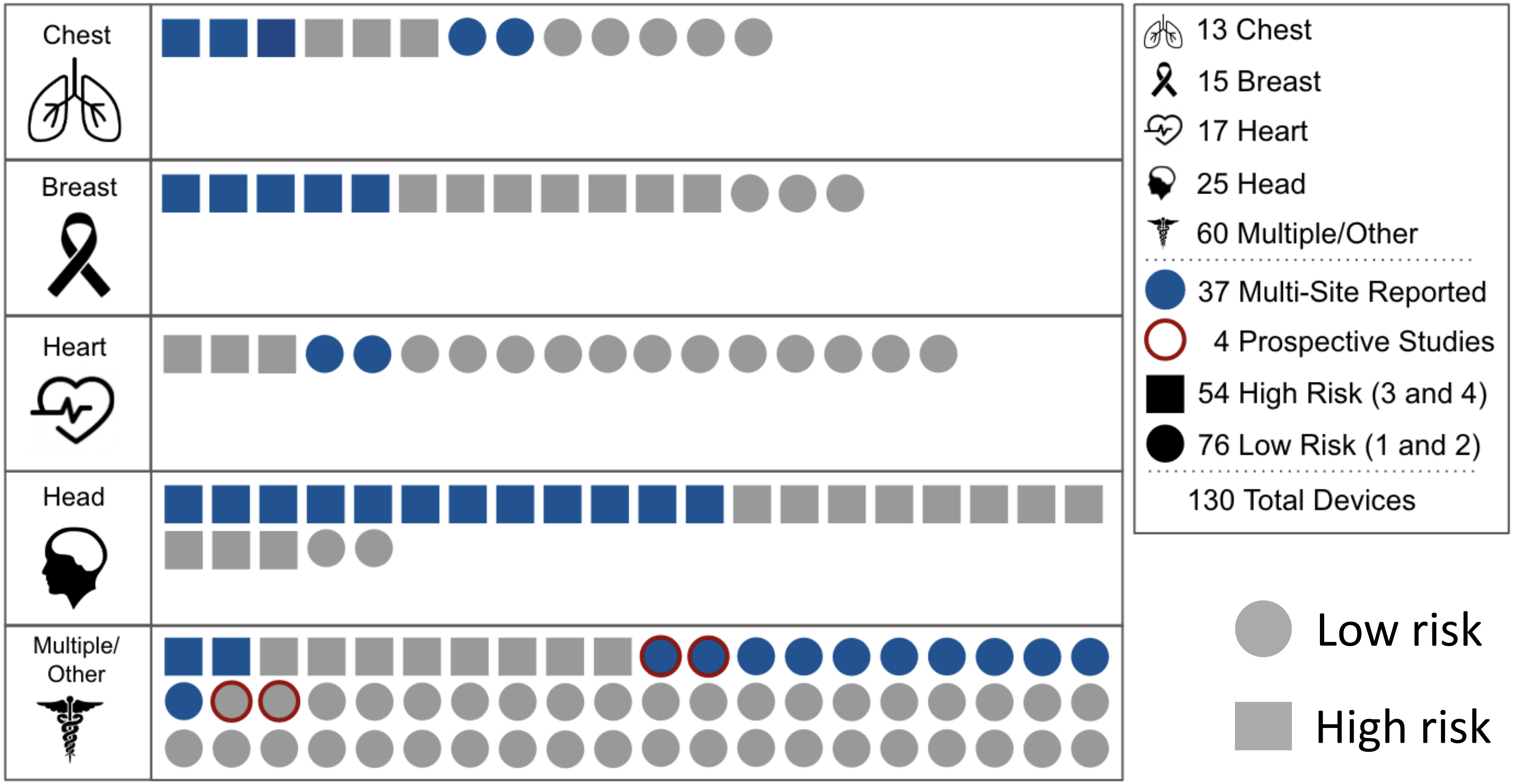
0.05 worse on darker skin

sensitivity: 0.41 (white) and 0.12 (dark)

few dark skin samples in orig. train/test

Roxana Daneshjou

Data used to test 130 FDA-approved AI



● Low risk

■ High risk

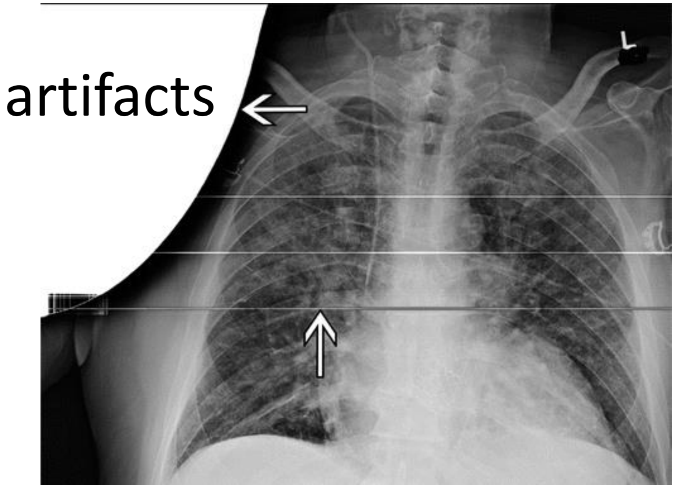
93/130 did not report multi-site evaluation

Only 4 prospective studies

Wu et al. *Nature Medicine* 2021

Large variability in cross site performance

Site	**Stanford** (N=19K)	**Boston** (N=23K)	**NIH** (N=11K)
Stanford	**0.90 ± 0.01**	0.87 ± 0.01	0.85 ± 0.02
Baylor	0.83 ± 0.01	**0.89 ± 0.01**	0.84 ± 0.02
NIH	0.78 ± 0.01	0.76 ± 0.02	**0.88 ± 0.02**



Pneumothorax detection

Wu et al. *Nature Medicine* 2021

Lessons for deploying trustworthy medical AI

1. Understand what data is used to develop the AI.

2. Understand why AI makes systematic mistakes.

3. Use human-in-the-loop evaluation.

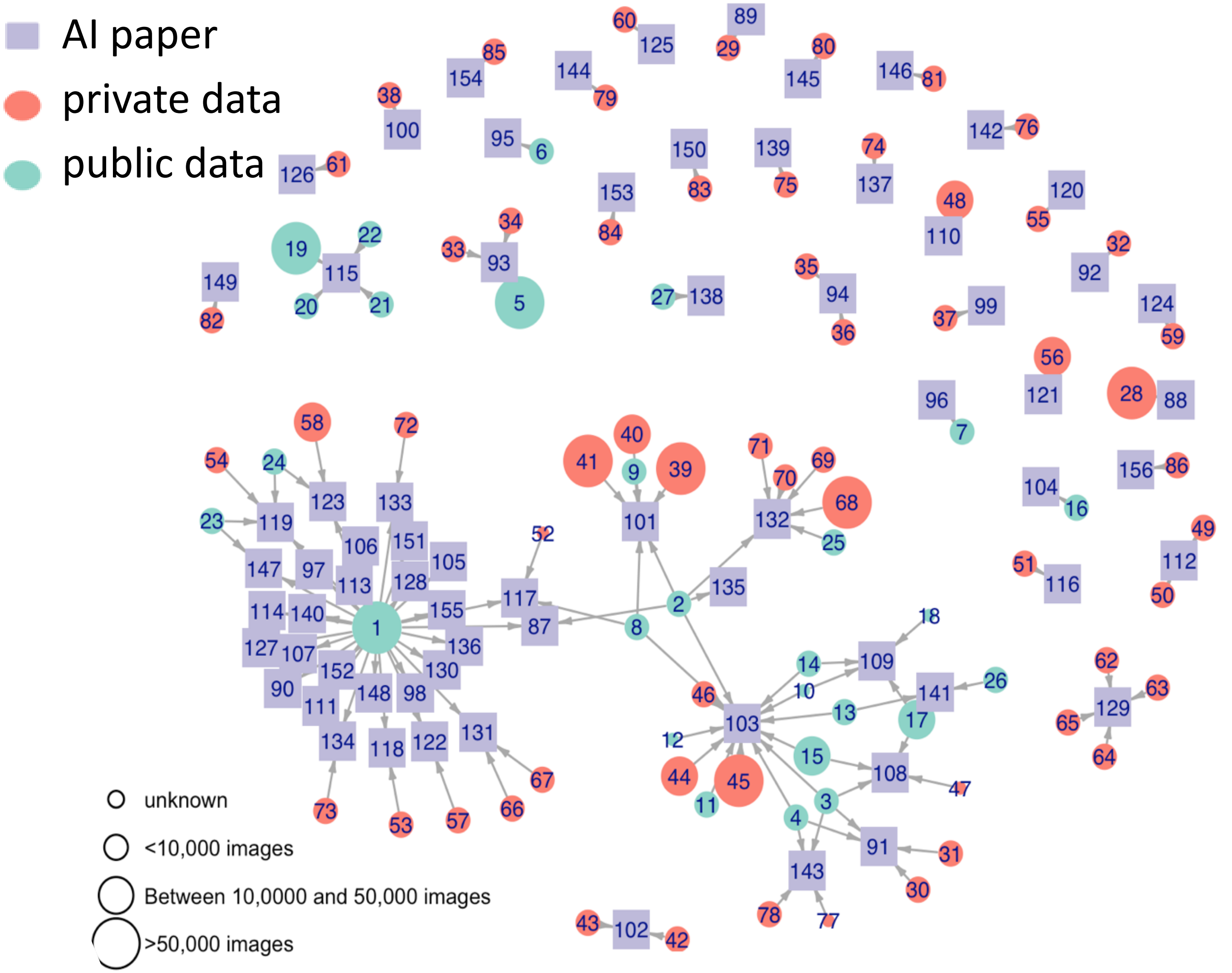
Lessons for deploying trustworthy medical AI

1. Understand what data is used to develop the AI.

2. Understand why AI makes systematic mistakes.

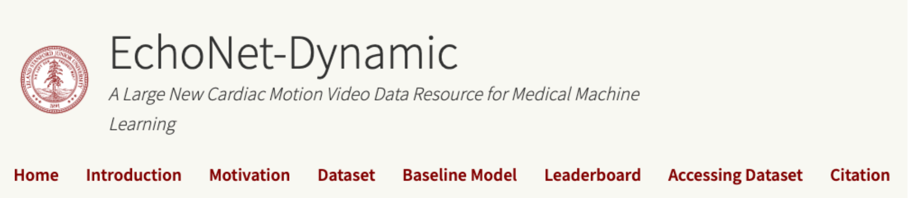
3. Use human-in-the-loop evaluation.

1. Data used to train dermatology AI



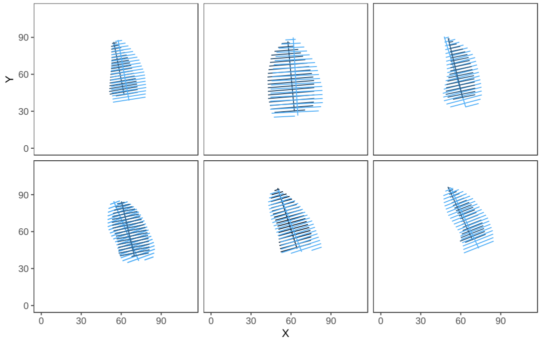
Daneshjou et al. *JAMA Derm* 202

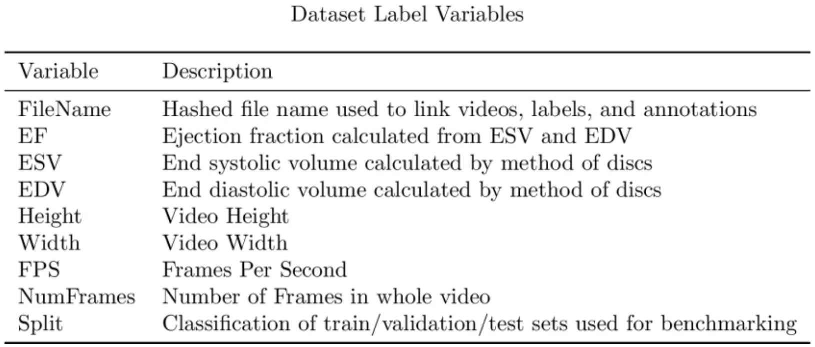
1. Transparent dataset and code



EchoNet-Dynamic

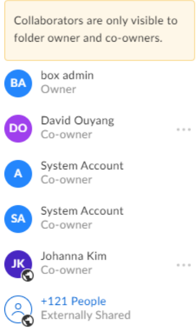
A Large New Cardiac Motion Video Data Resource for Medical Machine Learning





Dataset Label Variables

Variable	Description
FileName	Hashed file name used to link videos, labels, and annotations
EF	Ejection fraction calculated from ESV and EDV
ESV	End systolic volume calculated by method of discs
EDV	End diastolic volume calculated by method of discs
Height	Video Height
Width	Video Width
FPS	Frames Per Second
NumFrames	Number of Frames in whole video
Split	Classification of train/validation/test sets used for benchmarking



Collaborators are only visible to folder owner and co-owners.

BA	box admin	Owner
DO	David Ouyang	Co-owner
A	System Account	Co-owner
SA	System Account	Co-owner
JK	Johanna Kim	Co-owner
	+121 People	Externally Shared

Largest public dataset of medical videos.

Ouyang et al. *Nature* 2020

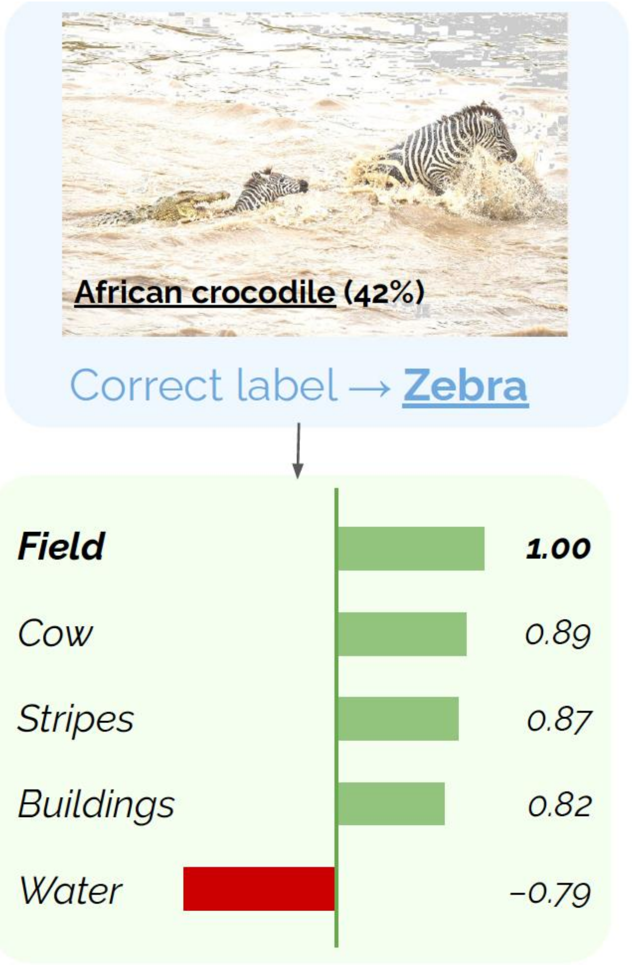
Lessons for deploying trustworthy medical AI

1. Understand what data is used to develop the AI.

2. Understand why AI makes systematic mistakes.

3. Use human-in-the-loop evaluation.

2. Why did the model make this mistake?

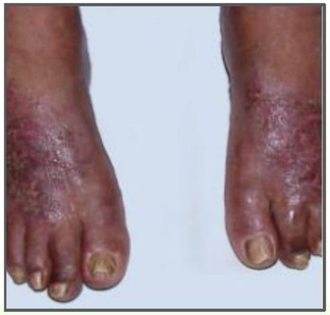


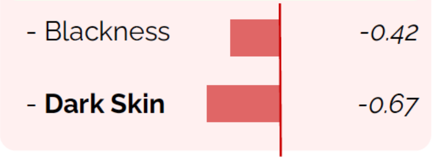
Conceptual explanation of mistakes

Conceptual explanation of mistakes

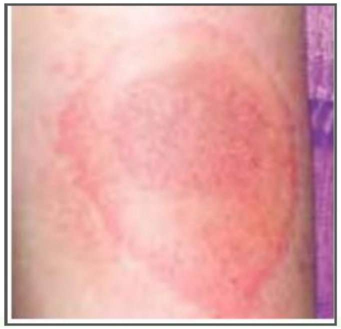
Mistakes made by the model

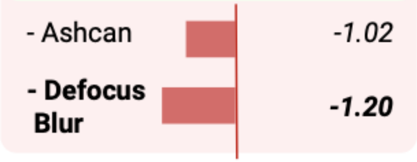
Label: Allergic Contact Dermatitis
Pred: Stasis Edema (19%)



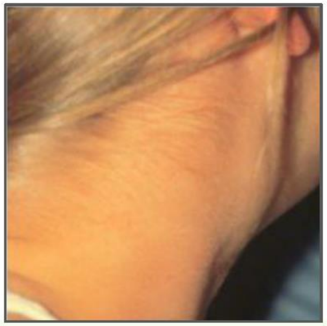


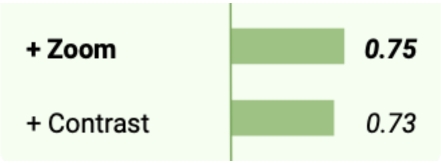
Label: Fixed Eruptions
Pred: Erythema Nodosum(35%)



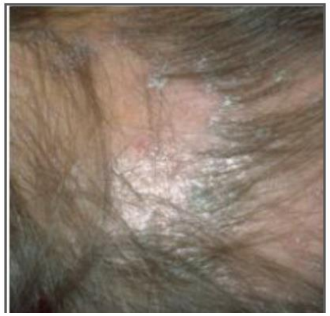


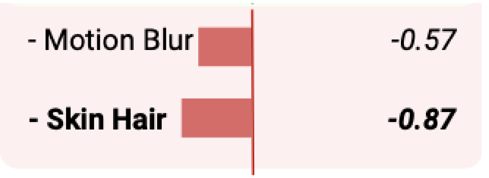
Label: Mucinosis
Pred: Aplasia Cutis (9%)





Label: Sarcoidosis
Pred: Nevus Sebaceous of Jadassohn (36%)





Output of our AI mistake explainer

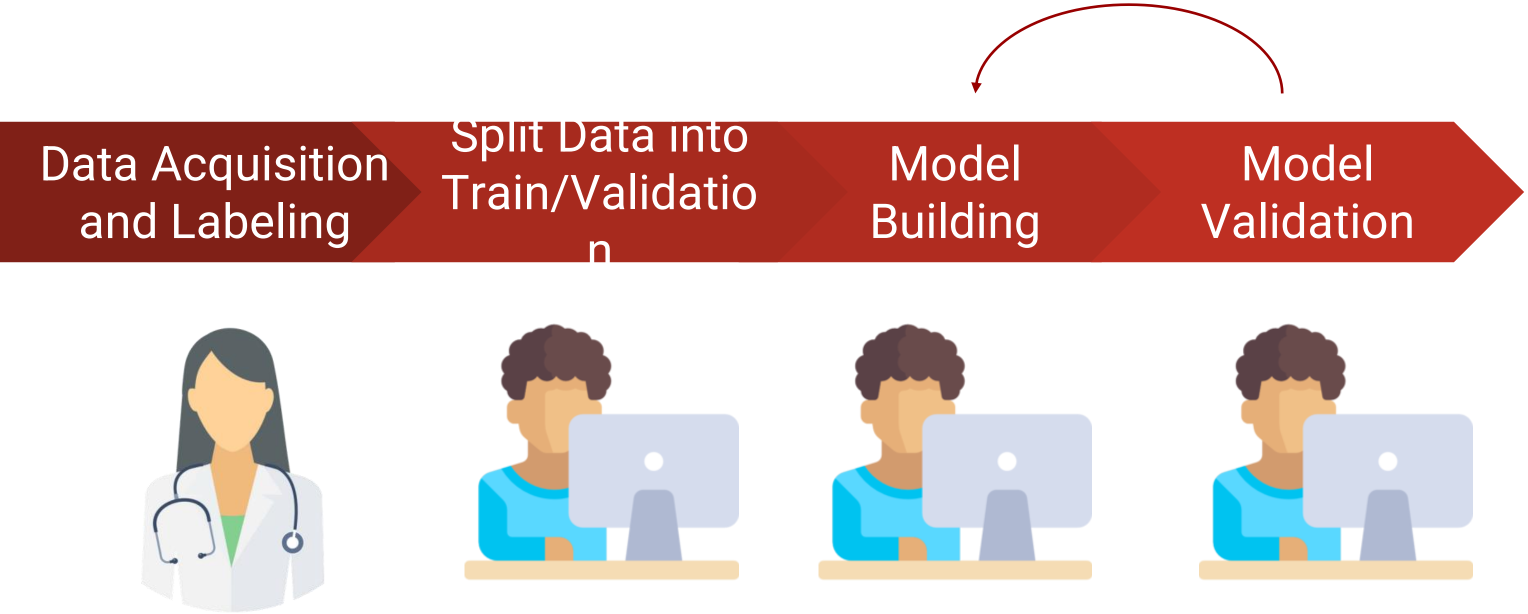
Abubakar Abid, Mert Yuksekgonul

Lessons for deploying trustworthy medical AI

1. Understand what data is used to develop the AI.
2. Understand why AI makes systematic mistakes.
3. Use human-in-the-loop evaluation.

3. AI often optimizes the wrong objective

optimization over a **fixed, single-site** validation dataset



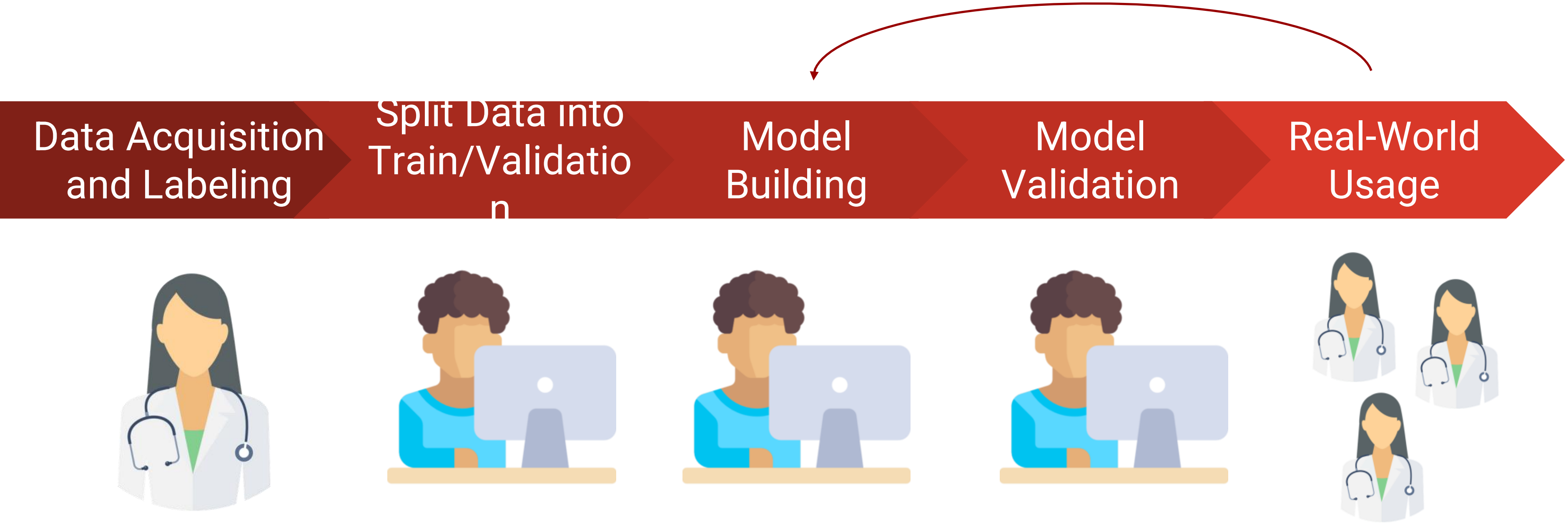
Data Acquisition and Labeling

Split Data into Train/Validatio
n

Model Building

Model Validation

Optimize for **human usage** instead!



optimize over **real-world user data**

Data Acquisition and Labeling

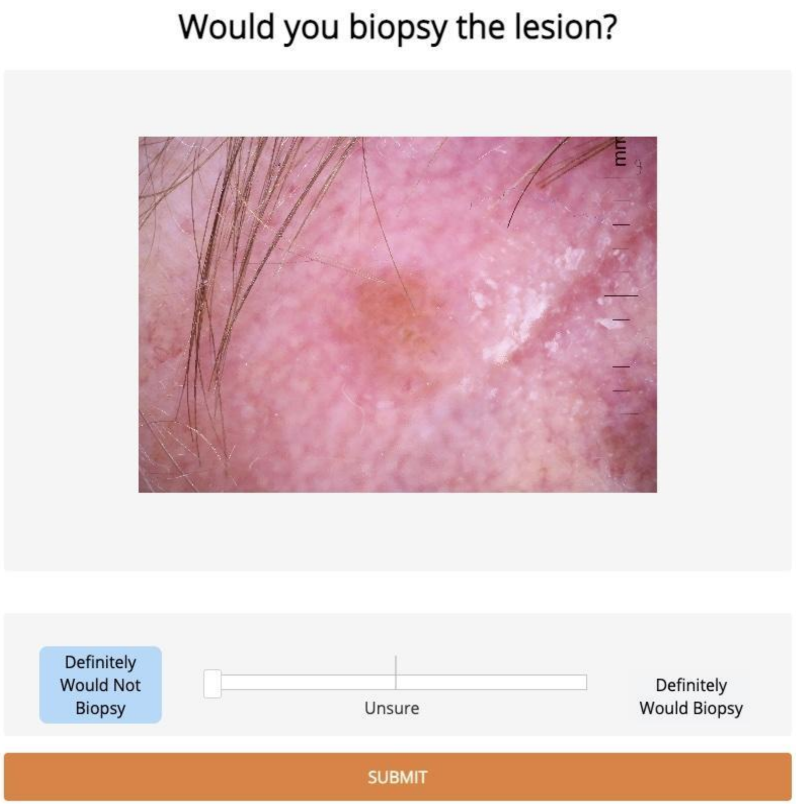
Split Data into Train/Validation

Model Building

Model Validation

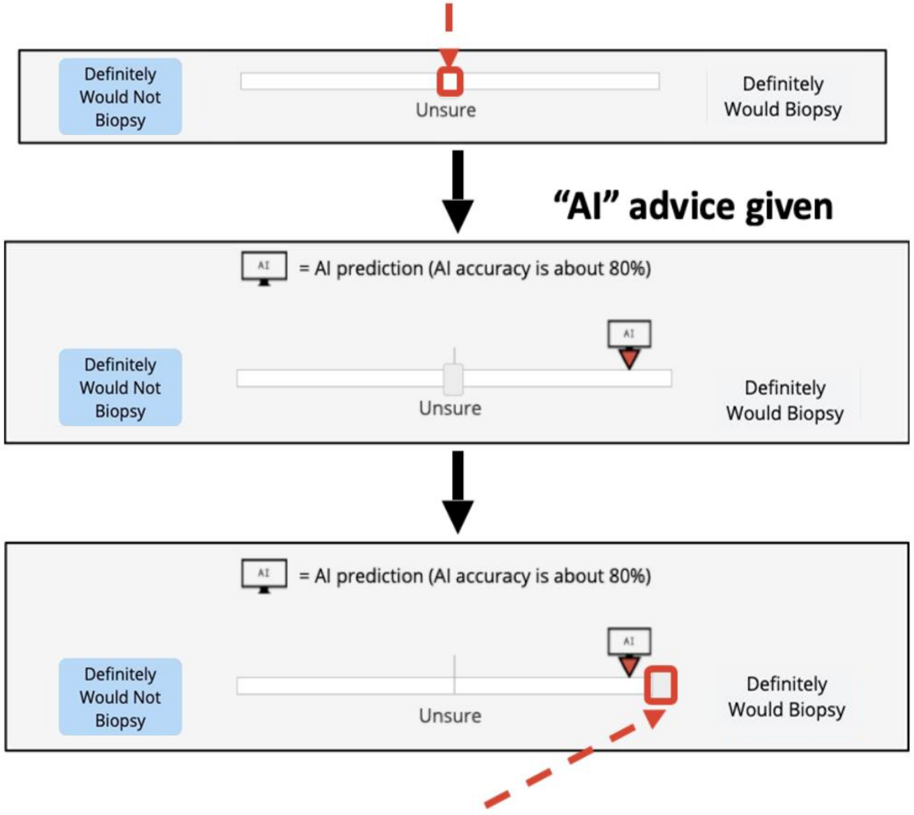
Real-World Usage

Human-in-the-loop evaluation of ML impact



Would you biopsy the lesion?

Participant's initial response (response 1)



"AI" advice given

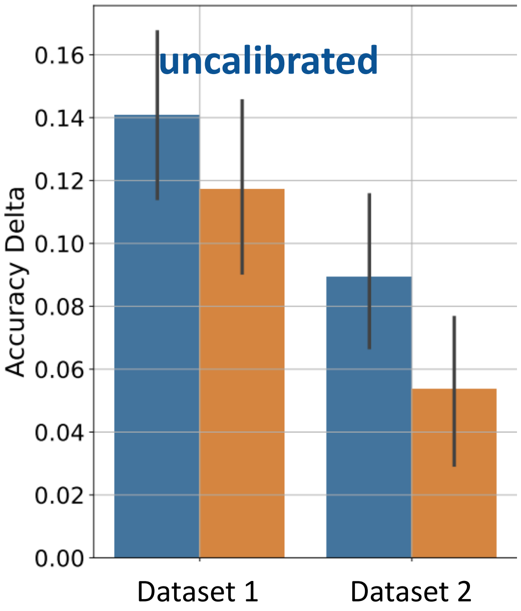
= AI prediction (AI accuracy is about 80%)

Participant's final response (response 2)

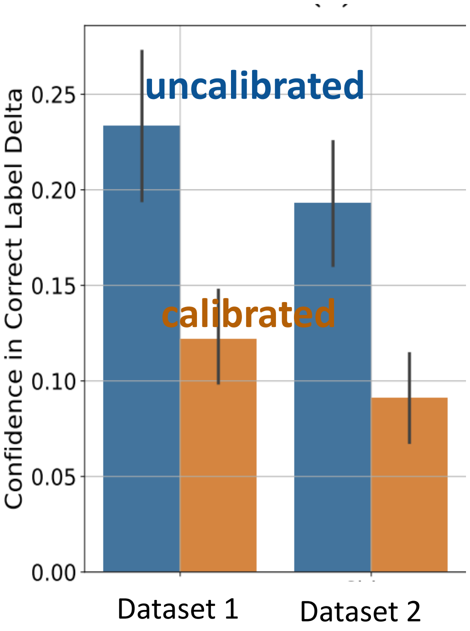
Kailas Vodrahalli

Worse AI can be better for humans

Human accuracy improvement



Human confidence in correct answer



Uncalibrated = overconfident models

Kailas Vodrahalli

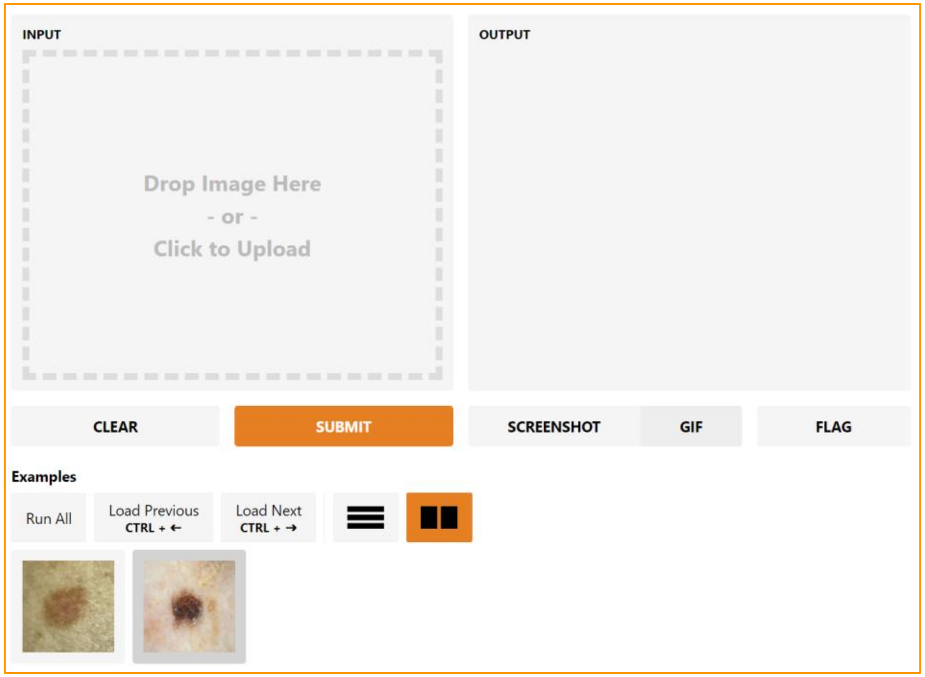
Using gradio

model to deploy

type of UI to create

```
import gradio
app = gradio.Interface(classify_skin_image, inputs=“image”, outputs=“label”)
app.launch(share=True)
```

url: www.gradio.app/test12543
can be shared



Abid et al *Nature Machine Intelligence* 202

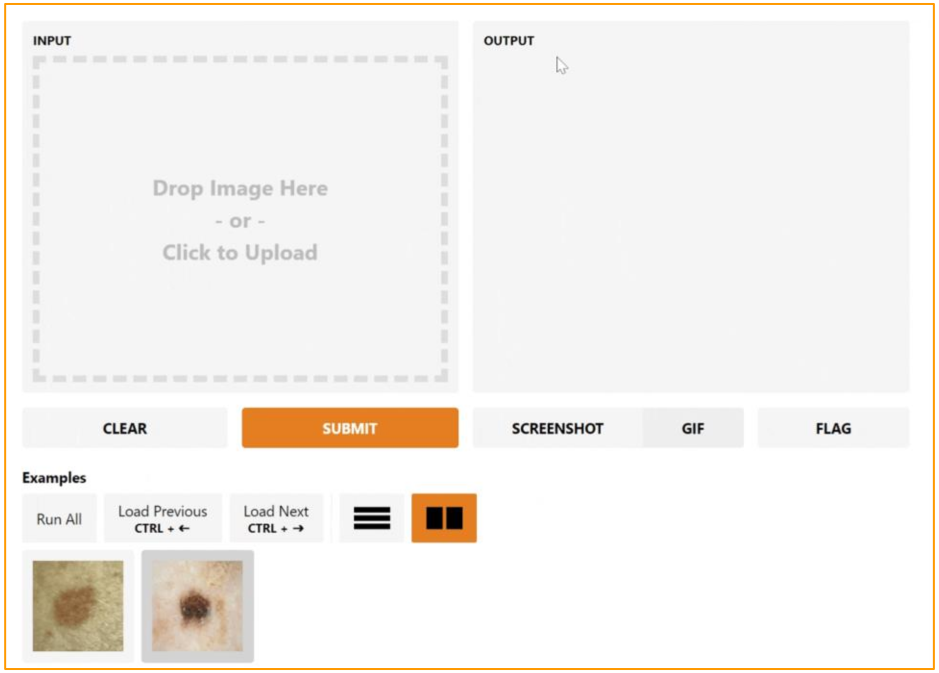
Using gradio

model to deploy

type of UI to create

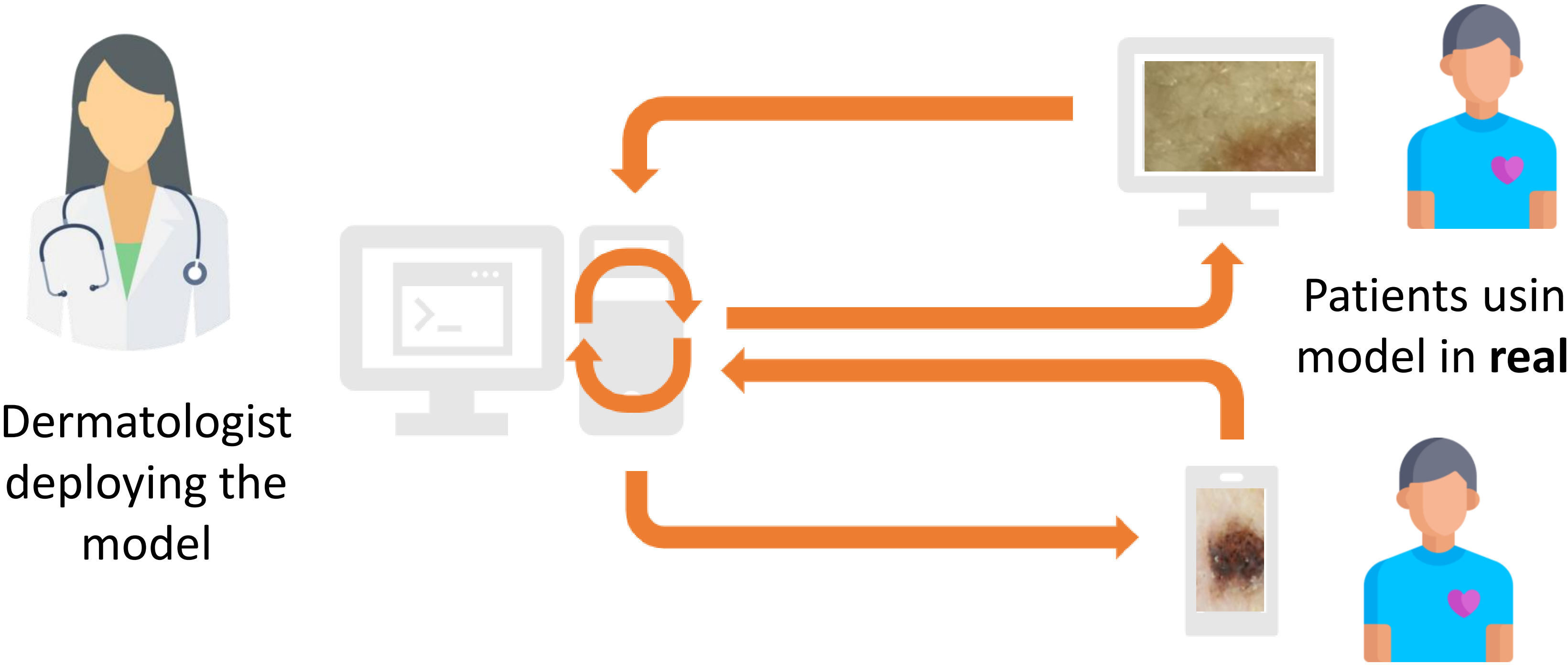
```
import gradio
app = gradio.Interface(classify_skin_image, inputs=“image”, outputs=“label”)
app.launch(share=True)
```

url: www.gradio.app/test12543
can be shared



Abid et al *Nature Machine Intelligence* 202

gradio used for Stanford’s 1st real-time AI trial



Dermatologist deploying the model

Patients using the model in **real time**

Google amazon facebook cisco

vmware SIEMENS MIT

UiPath™ UNIFYID HUMANISE FACTMATA WNS

Lessons for deploying trustworthy medical AI

1. Understand what data is used to develop the AI.

2. Understand why AI makes systematic mistakes.

3. Use human-in-the-loop evaluation.

Resources and thanks

Papers and codes available www.james-zou.com

Disparity in dermatology AI

Daneshjou et al. *JAMA Dermatology* 2021

Roxana Daneshjou

Data transparency for biomedical AI

Wu et al *Nature Medicine* 2021

Eric Wu

Video-based AI for cardiac assessment.

Ouyang et al. *Nature* 2020

David Ouyang

Explaining model mistakes

Abid, Yuksekgonul, Zou. In review 2022

Mert Yuksekgonal

Gradio for human-in-the-loop AI

Abid et al. *Nature Machine Intelligence* 2020

Abu Abid

Thanks to: NIH, NSF CAREER, Sloan, Chan-Zuckerberg