public datasets
X AI for children

October 16, 2026

Overview

1. children
2. public datasets
3. children in public datasets
4. AI for ~~children~~ adults

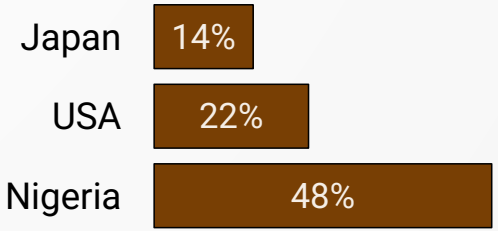
Overview

1. **children**
2. public datasets
3. children in public datasets
4. AI for ~~children~~ adults

Why talk about children?

Why talk about children?

1. make up ~30% of the global population



Why talk about children?

1. make up ~30% of the global population

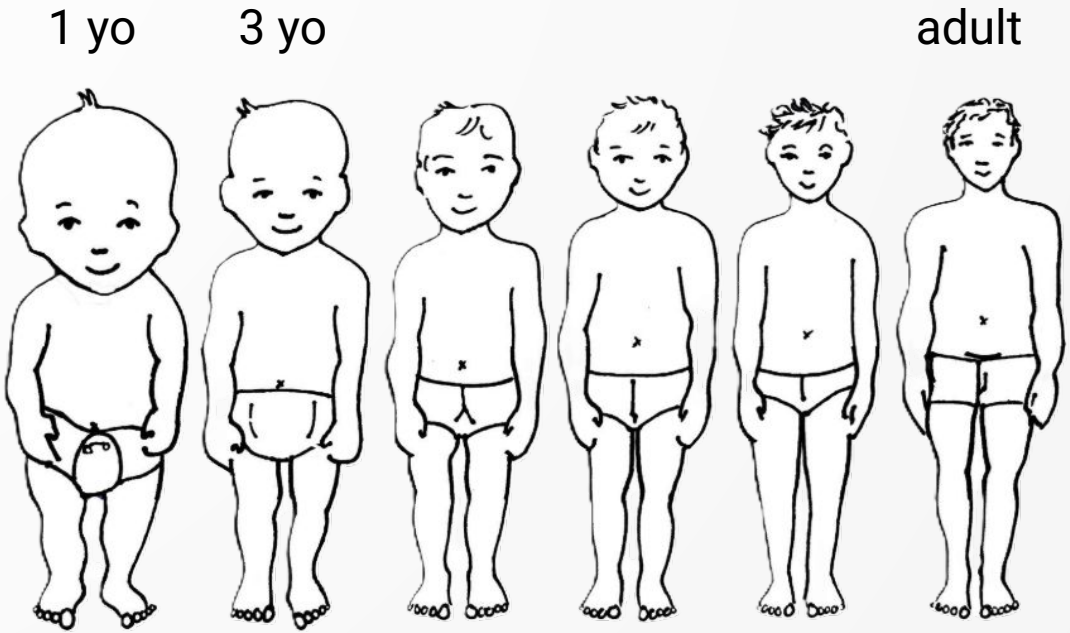
2. are vulnerable and rarely manage their own health

Why talk about **children**?

1. make up ~30% of the global population
2. are vulnerable and rarely manage their own health
3. “*children are not small adults*”

children are not small adults

children are not small adults in anatomy



anatomical proportions change over time

children are not small adults in anatomy physiology



immune system

7-8 years old



bones

16-21 years old



brain

25 years old

different body parts develop
at different ages

children are not small adults in anatomy

physiology

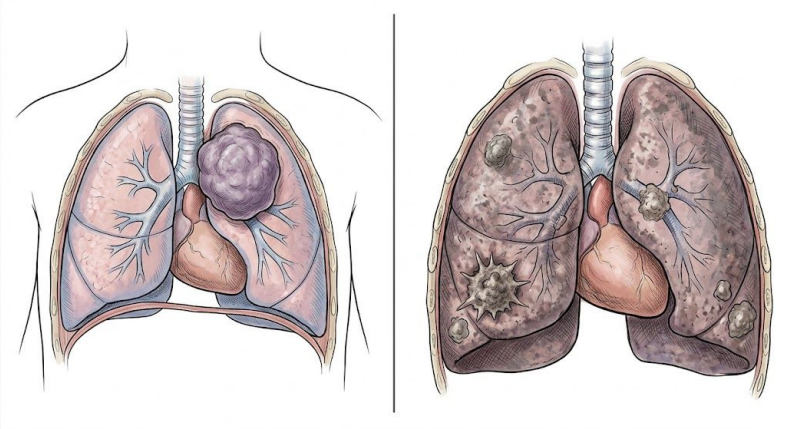
pathology

1. pediatric vs. adult diseases differ

(rare vs. chronic)

lung cancer

rare genetic mutation



long-term exposure to carcinogens

children are not small adults in anatomy

physiology

pathology

1. pediatric vs. adult diseases differ

   (rare vs. chronic)

2. children respond differently to treatments

children are not small adults in anatomy

physiology

pathology

1. pediatric vs. adult diseases differ

   (rare vs. chronic)

2. children respond differently to treatments

3. children have a harder time communicating their symptoms



children are a **very diverse** population

Overview

1. **children**
2. public datasets
3. children in public datasets
4. AI for ~~children~~ adults

Overview

1. children
2. **public datasets**
3. children in public datasets
4. AI for ~~children~~ adults

Why talk about **public datasets**?

(for machine learning)

public datasets play four roles

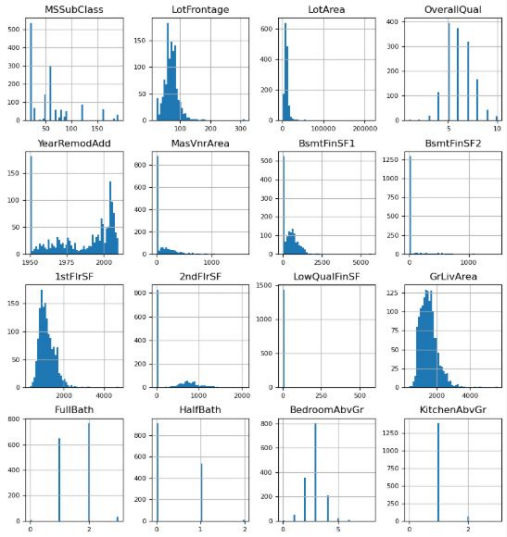
1. educate
2. draw attention to important problems
3. support model development
4. assist in model evaluation

public datasets educate

KAGGLE · GETTING STARTED PREDICTION COMPETITION · ONGOING

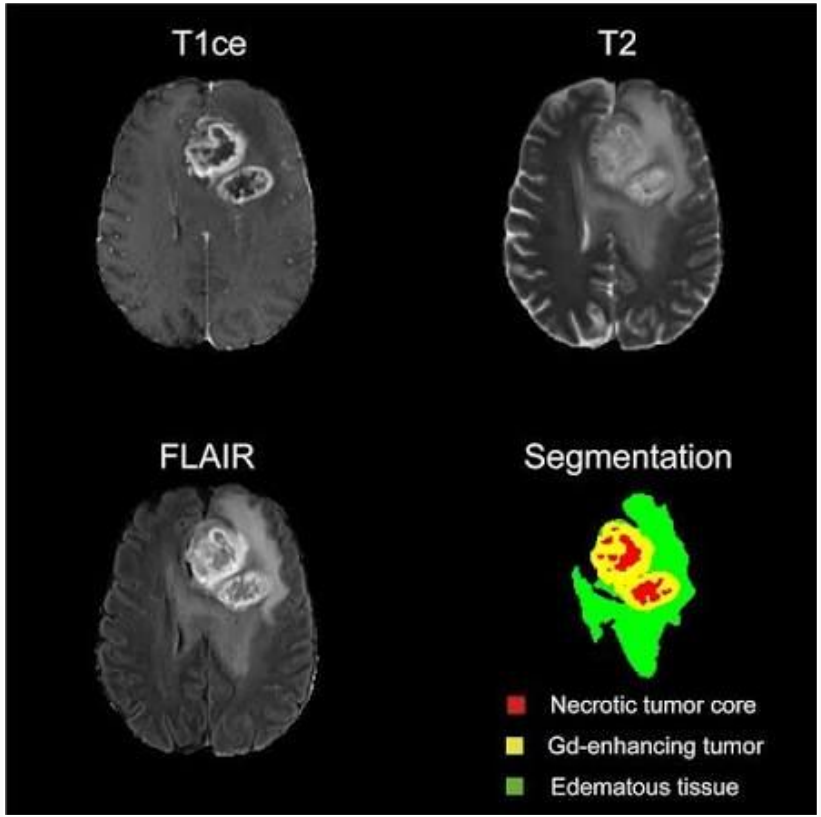
House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting



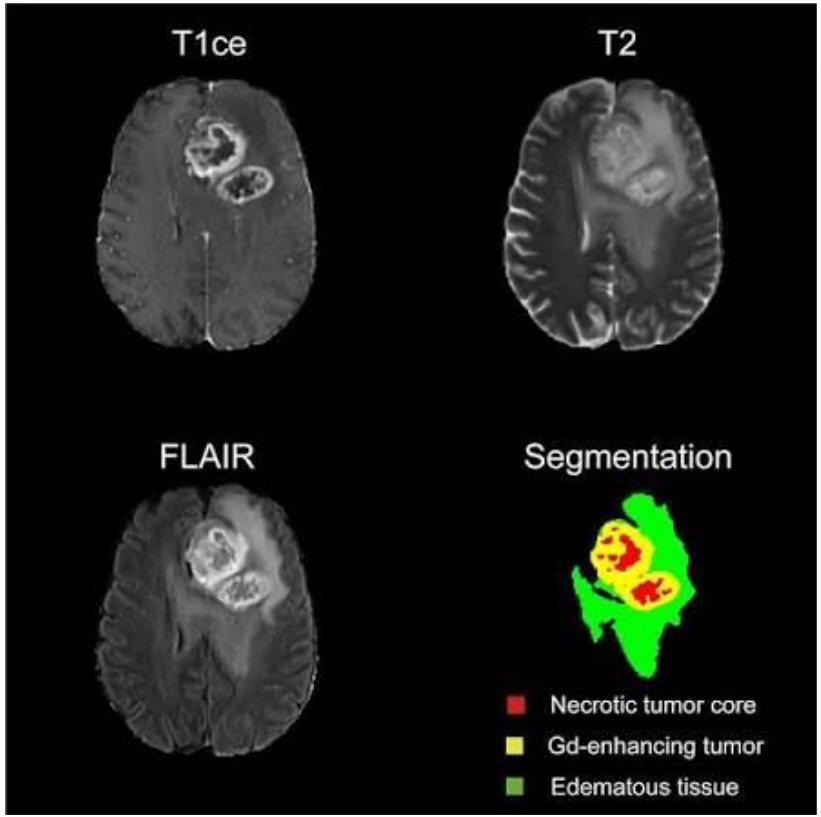

```
from sklearn.ensemble import RandomForestRegressor
...
model = RandomForestRegressor()
model.fit(X_train, y_train)
y_test_pred = model.predict(X_test)
```

public datasets draw attention to important problems



Brain Tumor Segmentation (BraTS)
MICCAI Challenge (2012 to 2026)

public datasets draw attention to important problems

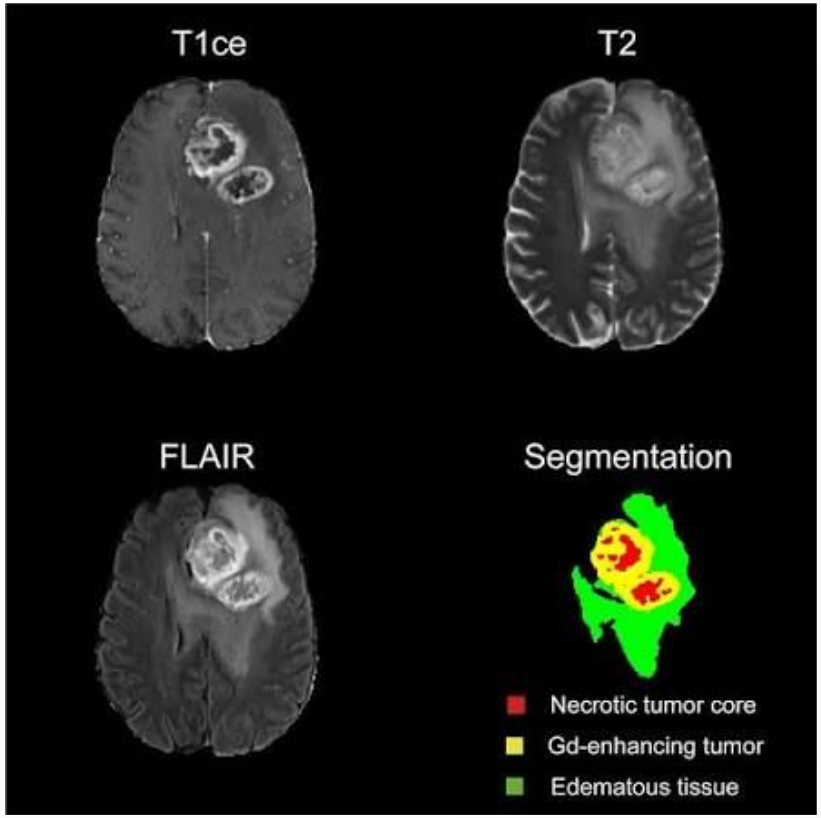


Brain Tumor Segmentation (BraTS)
MICCAI Challenge (2012 to 2026)

given multiple MRI scans,
segment three tumor subregions:

- necrotic tumor core (dead)
- contrast-enhanced tumor (leaky)
- edematous (inflamed)

public datasets draw attention to important problems



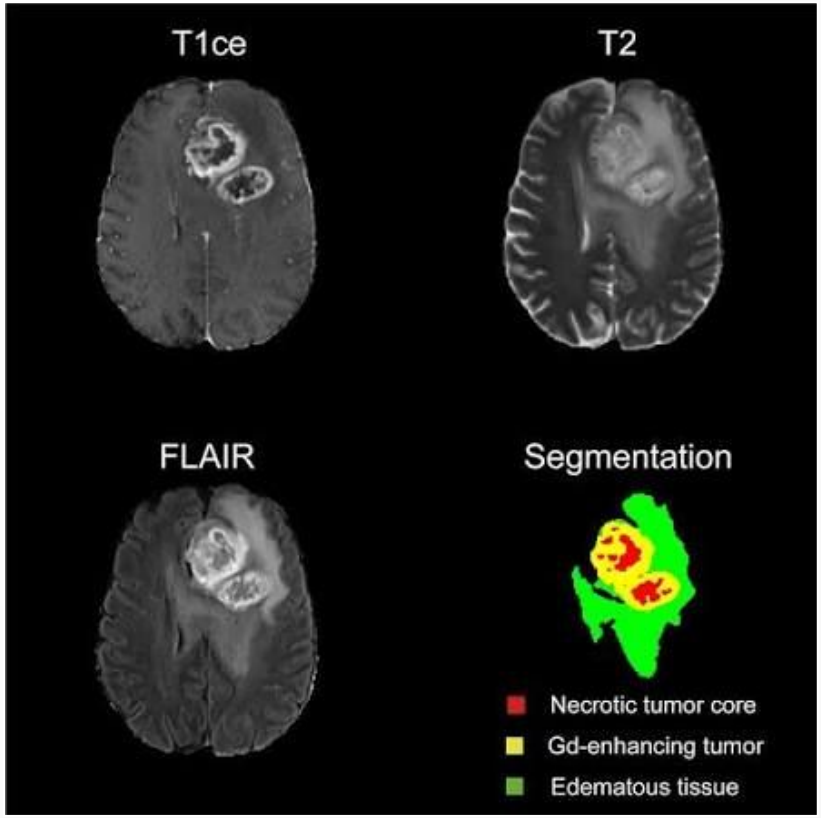
Brain Tumor Segmentation (BraTS)
MICCAI Challenge (2012 to 2026)

given multiple MRI scans,
segment three tumor subregions:

- necrotic tumor core (dead)
- contrast-enhanced tumor (leaky)
- edematous (inflamed)

became one of the most important benchmarks for segmentation models

public datasets support model development



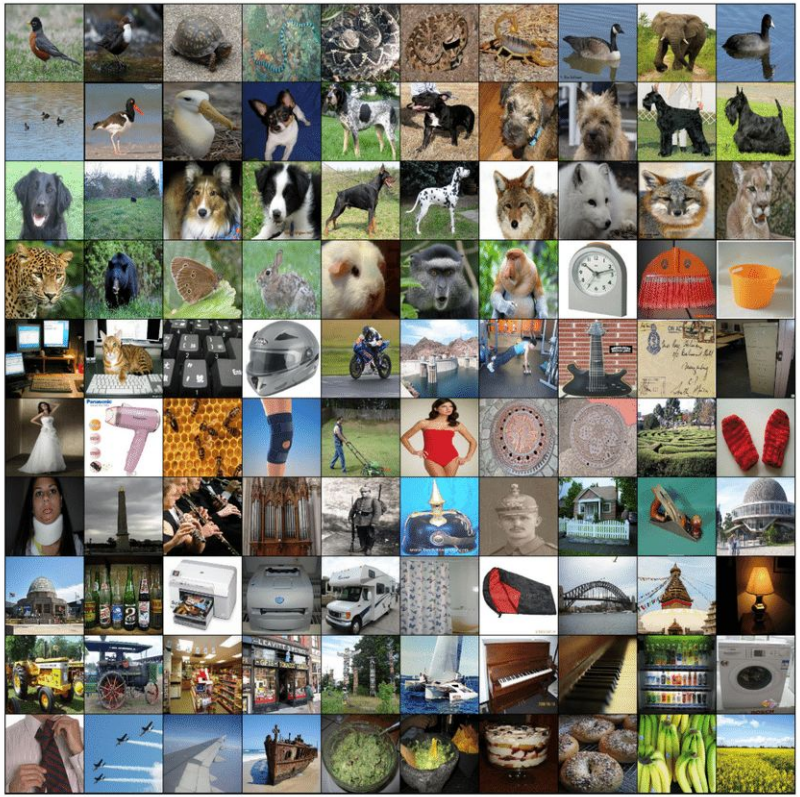
Brain Tumor Segmentation (BraTS)
MICCAI Challenge (2012 to 2026)

given multiple MRI scans,
segment three tumor subregions:

- necrotic tumor core (dead)
- contrast-enhanced tumor (leaky)
- edematous (inflamed)

became one of the most important benchmarks for segmentation models

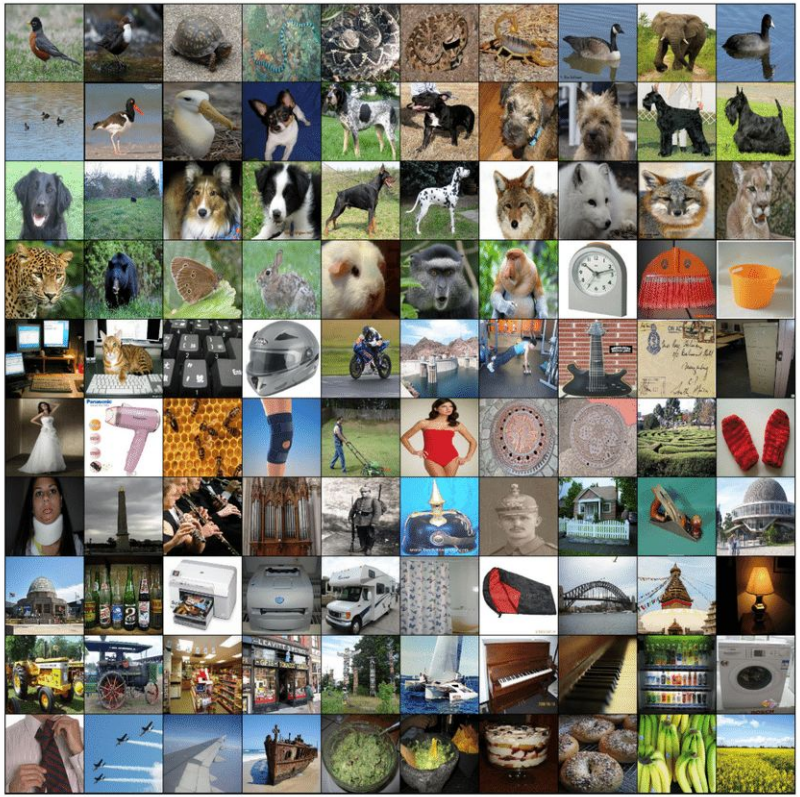
public datasets support model development



ImageNet-1K dataset

given image, classify with 1 of 1000 labels

public datasets support model development



ImageNet-1K dataset

given image, classify with 1 of 1000 labels

enabled many advances in ML

1. architectures (ResNet)
2. data augmentations (MixUp)
3. transfer learning
4. self-supervised learning (SimCLR)
5. few-shot learning
6. knowledge distillation
7. distribution shift
8. outlier detection

public datasets assist in model evaluation

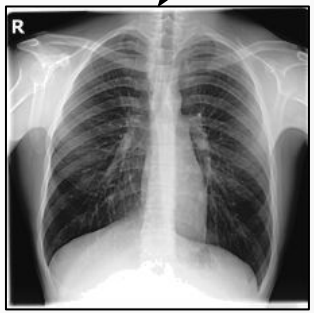


VinDr-CXR
Hanoi, Vietnam

Acc: 98%



model



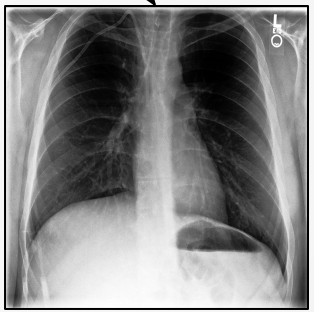
PadChest
Barcelona, Spain

Acc: 87%



NIH X-ray14
USA (NIH)

Acc: 96%



CheXBERT
USA (Stanford)

Acc: 77%

What is one consequence of public datasets?

public datasets make it easy to forget the **human**

1. **who** decided what data to collect?

2. **who** collected the data?

3. **who** annotated the data?

4. **whose** data is in the dataset?



VinDr-CXR

Hanoi, Vietnam

1. **who** decided what data to collect?

researchers
(funding agencies and ethics board by proxy)



VinDr-CXR
Hanoi, Vietnam



VinDr-CXR
Hanoi, Vietnam

1. **who** decided what data to collect?

researchers
(funding agencies and ethics board by proxy)

2. **who** collected the data?

X-ray technicians in the Hanoi hospital



VinDr-CXR
Hanoi, Vietnam

1. **who** decided what data to collect?

researchers
(funding agencies and ethics board by proxy)

2. **who** collected the data?

X-ray technicians in the Hanoi hospital

3. **who** annotated the data?

radiologist in the Hanoi hospital



VinDr-CXR
Hanoi, Vietnam

1. **who** decided what data to collect?

researchers
(funding agencies and ethics board by proxy)

2. **who** collected the data?

X-ray technicians in the Hanoi hospital

3. **who** annotated the data?

radiologist in the Hanoi hospital

4. **whose** data is in the dataset?

Vietnamese, 18+ years old,
who received a chest X-ray in the Hanoi hospital



VinDr-CXR
Hanoi, Vietnam

1. **who** decided what data to collect?

researchers
(funding agencies and ethics board by proxy)

2. **who** collected the data?

X-ray technicians in the Hanoi hospital

3. **who** annotated the data?

radiologist in the Hanoi hospital

4. **whose** data is in the dataset?

Vietnamese, **18+ years old**,
who received a chest X-ray in the Hanoi hospital

Overview

1. children
2. **public datasets**
3. children in public datasets
4. AI for ~~children~~ adults

Overview

1. children
2. public datasets
3. **children in public datasets**
4. AI for ~~children~~ adults

children

1. make up ~30% of the population
2. *are not small adults*

public datasets

1. educate
2. draw attention to important problems
3. support model development
4. assist in model evaluation

children

1. make up ~30% of the population
2. *are not small adults*

public datasets

1. educate
2. draw attention to important problems
3. support model development
4. assist in model evaluation

RQ: Are **children** well-represented in **public medical imaging datasets**?

RQ: Are **children** well-represented in **public medical imaging datasets**?

we reviewed **203 public datasets** of X-ray, CT, MRI, ultrasound and fundus images

Challenges (n=58)

1. Grand Challenge (44)
2. Kaggle (14)

Benchmarks (n=18)

1. MSD (10)
2. MedFAIR (7)
3. AMOS (1)

Collections (n=106)

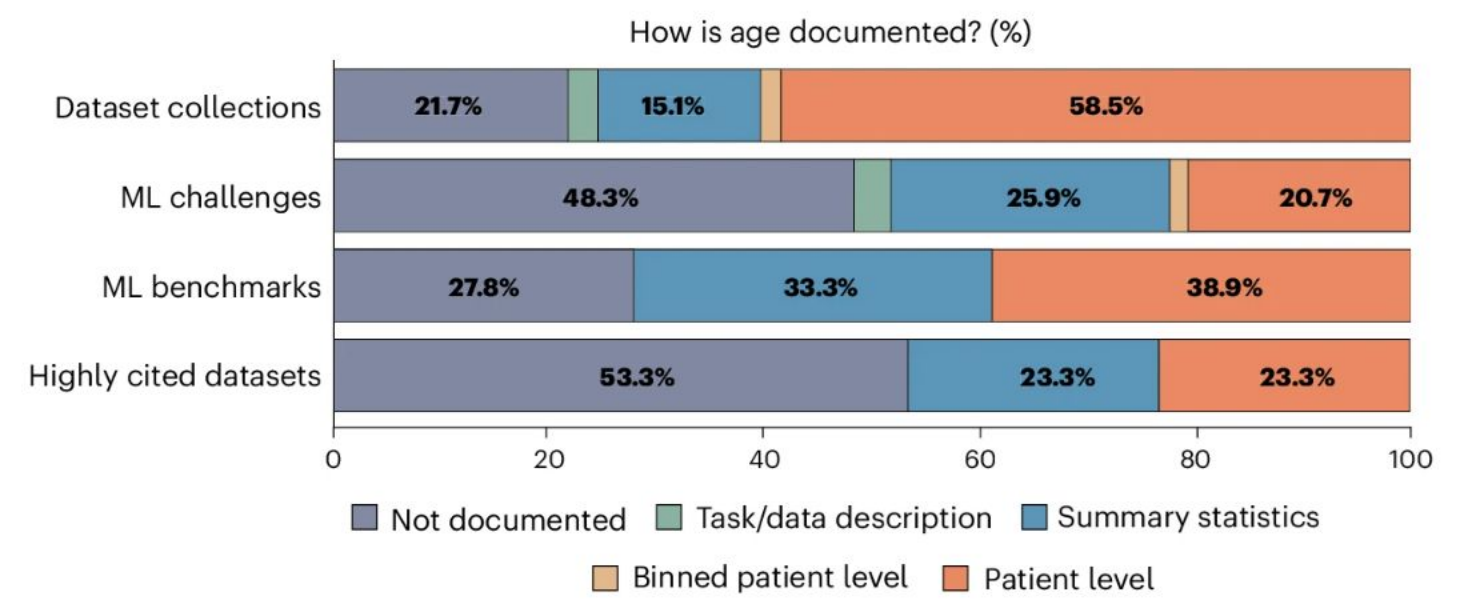
1. TCIA (69)
2. Stanford AIMI (13)
3. MIDRC (24)

Highly-Cited (n=21)

1. PapersWithCode Top-Cited

RQ: Are **children** well-represented in **public medical imaging datasets**?

we reviewed **203 public datasets** of X-ray, CT, MRI, ultrasound and fundus images



Finding 1. **many datasets do not report patient ages**
(so we estimate the # of children)

RQ: Are children well-represented in **public medical imaging datasets**?

we reviewed ~~203~~ **135** **public datasets** of X-ray, CT, MRI, ultrasound and fundus images

Challenges (n=~~58~~ 29)

1. Grand Challenge (~~44~~ 24)
2. Kaggle (~~14~~ 6)

Benchmarks (n=~~18~~ 15)

1. MSD (~~10~~ 7)
2. MedFAIR (~~7~~ 7)
3. AMOS (~~1~~ 1)

Collections (n=~~106~~ 80)

1. TCIA (~~69~~ 51)
2. Stanford AIMI (~~13~~ 10)
3. MIDRC (~~24~~ 19)

Highly-Cited (n=~~21~~ 11)

1. PapersWithCode Top-Cited

Finding 1. many datasets do not report patient ages

(so we estimate the # of children)

RQ: Are children well-represented in **public medical imaging datasets**?

we reviewed ~~203~~ 135 **public datasets** of X-ray, CT, MRI, ultrasound and fundus images

Challenges (n=~~58~~ 29)

Patients: 101,181
Children: 212

Benchmarks (n=~~18~~ 15)

Patients: 142,721
Children: 83

Collections (n=~~106~~ 80)

Patients: 231,077
Children: 4,043

Highly-Cited (n=~~21~~ 11)

Patients: 37,629
Children: 1,688

Main Finding. children represent under 2% of all patients (1.18%)

Patients: **512,608** # Children: **6,026**

RQ: Are **children** well-represented in **public medical imaging datasets**?

Main Finding. children represent under 2% of all patients (1.18%)

Patients: **512,608** # Children: **6,026**

RQ: Are **children** well-represented in **public medical imaging datasets**?

Main Finding. children represent under 2% of all patients (1.18%)

Patients: **512,608** # Children: **6,026**

Is this inflated by **cancer** datasets? **no** (1.27%)

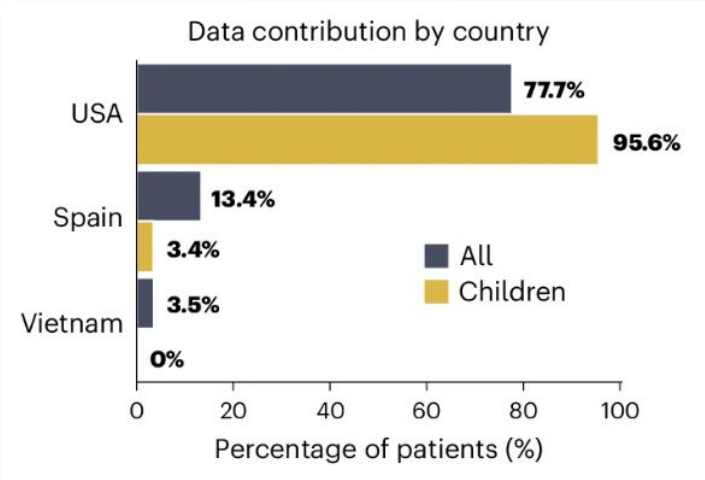
Patients: **463,180** # Children: **5,883**

RQ: Are **children** well-represented in **public medical imaging datasets**?

Main Finding. **children** represent **under 2%** of all patients (1.18%)

Patients: **512,608** # Children: **6,026**

Where do these patients come from? the **US** (largely)



RQ: Are **children** well-represented in **public medical imaging datasets**?

Main Finding. **children** represent **under 2%** of all patients (1.18%)

Patients: **512,608** # Children: **6,026**

Does this vary by **imaging modality** and **prediction task**?

RQ: Are children well-represented in **public medical imaging datasets**?

Main Finding. children represent under 2% of all patients (1.18%)

Patients: **512,608** # Children: **6,026**

Does this vary by **imaging modality** and **prediction task**?

yes **yes**

RQ: Are **children** well-represented in **public medical imaging datasets**?

Main Finding. children represent under 2% of all patients (1.18%)

Patients: **512,608** # Children: **6,026**

Does this vary by **imaging modality** and **prediction task**?

yes

	All	
	Adult	*Peds*
X-ray	959,482	7,520
US	23,767	4,467
CT	289,581	914
MRI	52,390	176
Fundus	117,900	48

images

X-ray (128 : 1)
US (6 : 1)
CT (317 : 1)
MRI (298 : 1)
Fundus (2457 : 1)

RQ: Are **children** well-represented in **public medical imaging datasets**?

Main Finding. children represent under 2% of all patients (1.18%)

Patients: **512,608** # Children: **6,026**

Does this vary by **imaging modality** and **prediction task**?

yes

	All	
	Adult	*Peds*
X-ray	959,482	7,520
US	23,767	4,467
CT	289,581	914
MRI	52,390	176
Fundus	117,900	48

images

X-ray (128 : 1)
US (6 : 1)
CT (317 : 1)
MRI (298 : 1)
Fundus (2457 : 1)

"6 adult ultrasound images for every 1 pediatric"

RQ: Are **children** well-represented in **public medical imaging datasets**?

Main Finding. **children** represent **under 2%** of all patients (1.18%)

Patients: **512,608** # Children: **6,026**

Does this vary by **imaging modality** and **prediction task**?

yes

	All	
	Adult	*Peds*
X-ray	959,482	7,520
US	23,767	4,467
CT	289,581	914
MRI	52,390	176
Fundus	117,900	48

images

X-ray (128 : 1)
US (6 : 1)
CT (317 : 1)
MRI (298 : 1)
Fundus (2457 : 1)

"300 adult CT/MRI images for every 1 pediatric"

RQ: Are **children** well-represented in **public medical imaging datasets**?

Main Finding. **children** represent **under 2%** of all patients (1.18%)

Patients: **512,608** # Children: **6,026**

Does this vary by **imaging modality** and **prediction task**?

yes

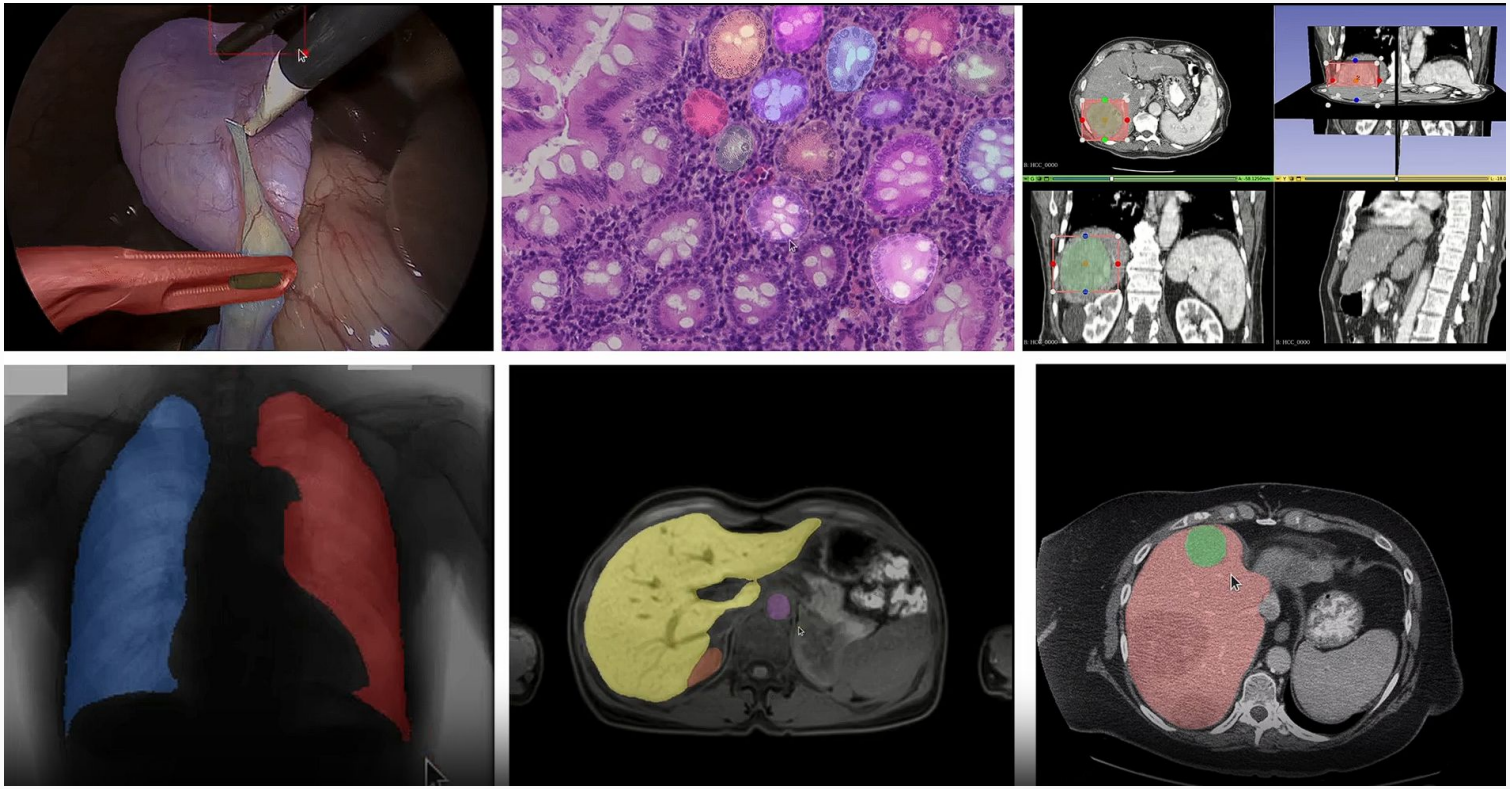
	Condition classification		Anatomy segmentation		Lesion segmentation		Image enhancement	
	Adult	*Peds*	*Adult*	*Peds*	*Adult*	*Peds*	*Adult*	*Peds*
X-ray	891,720	6136						
US	192	0	10,830	4,467	751	0	518	0
CT	262,069	124	4,677	373	7,092	309	760	40
MRI	18,905	5	1,903	17	22,174	134	1,283	4
Fundus	117,860	20	40	28				

Why should we care?

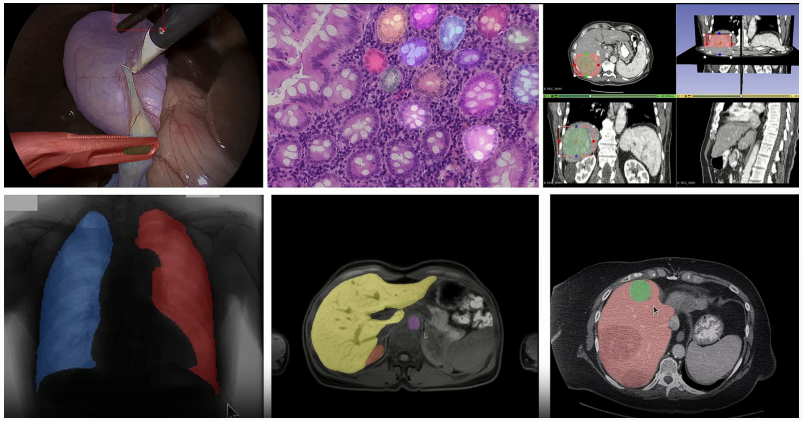
few **children** in **public datasets** means

1. ~~**educate on differences** between adults and children~~
2. ~~**draw attention** to problems specific to children's health~~
3. ~~support **model development** for children~~
4. ~~assist in **model evaluation** for children~~

Example. **MedSAM** (a foundation model for medical image segmentation)



Example. **MedSAM** (a foundation model for medical image segmentation)



1.6M CT images, **8508** patients

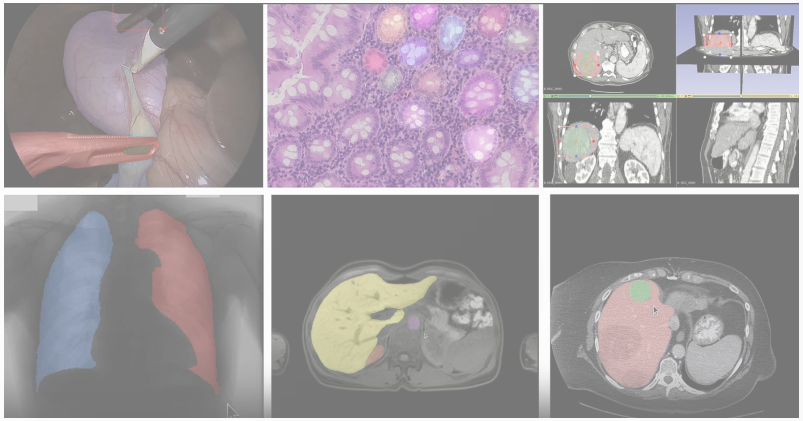
26 public CT datasets

→ **11** don't report age

→ **12** are adult-only

→ **3** contain adults and children

Example. **MedSAM** (a foundation model for medical image segmentation)



1.6M CT images, **8508** patients

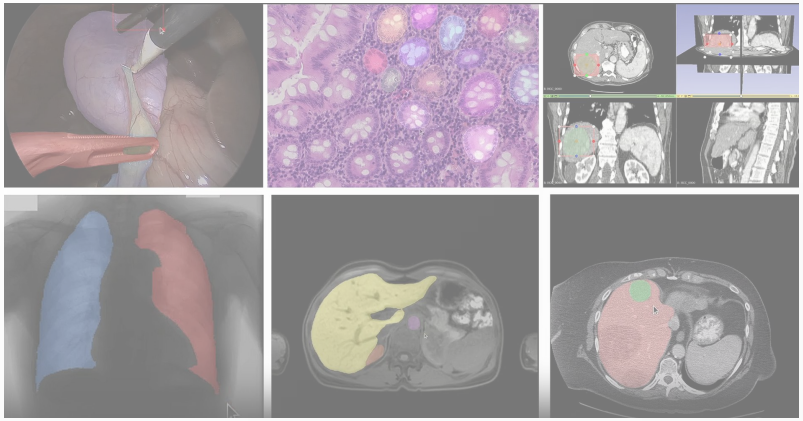
26 public CT datasets

- → **11** don't report age
- → **12** are adult-only
- → **3** contain adults and children

How many of the 8508 patients are children?

a) 0 **b)** 8 (0.1%) **c)** 85 (1%) **d)** 170 (2%) **e)** 8508 (100 %)

Example. **MedSAM** (a foundation model for medical image segmentation)



1.6M CT images, **8508** patients

26 public CT datasets

- → **11** don't report age
- → **12** are adult-only
- → **3** contain adults and children

How many of the 8508 patients are children?

14 (0.2%)



Overview

1. children
2. public datasets
3. **children in public datasets**
4. AI for ~~children~~ adults

Overview

1. children
2. public datasets
3. children in public datasets
4. **AI for ~~children~~ adults**

medicine has a bad history of treating children with adult interventions

1. **a common drug** (used to treat headaches) can cause brain / liver damage
2. **adult catheters** (inserted into veins) can cause bleeding and infection

medicine has a bad history of treating children with adult interventions

1. **a common drug** (used to treat headaches) can cause brain / liver damage
2. **adult catheters** (inserted into veins) can cause bleeding and infection

this happens because of **a lack of trial evidence** and **regulation**

medicine has a bad history of treating children with adult interventions

1. **a common drug** (used to treat headaches) can cause brain / liver damage
2. **adult catheters** (inserted into veins) can cause bleeding and infection

this happens because of **a lack of trial evidence** and **regulation**

Q: Are we making the same mistake with AI?

Approvals from the **Food and Drug Administration** (FDA) reflect a **lack of rigor** in evaluating **AI for children**

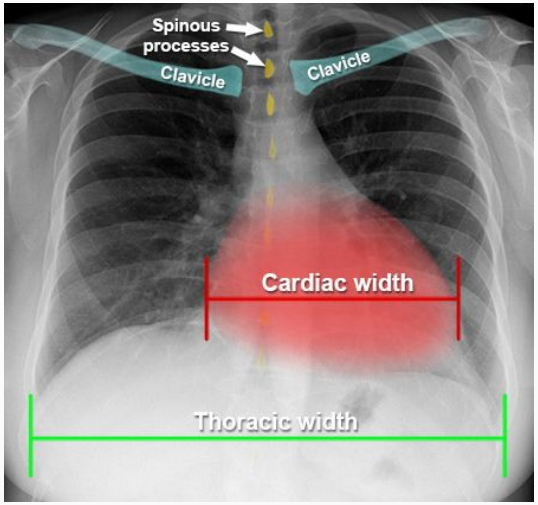
between 1995 to 2024,

- → **876** AI/ML devices were approved in total
  - → 639 (73%) omitted age information in validation studies
  - → **149** were authorized for pediatric use
    - → **28** reported validation on children
    - → **22** reported validation only on adults

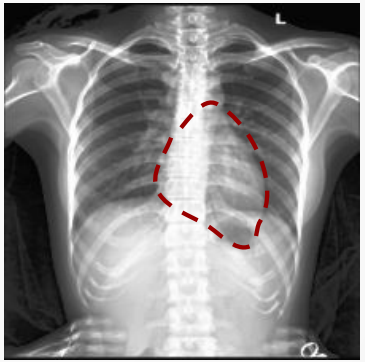
Would adult AI generalizes to children anyways?

RQ: Would adult AI generalizes to children anyways?

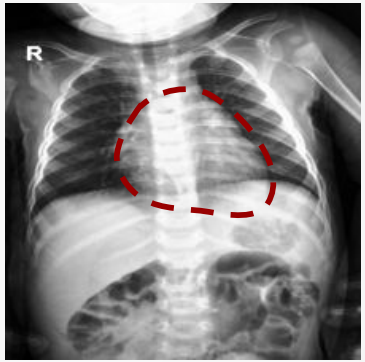
Case study. **Cardiomegaly** classification



enlarged heart



23 years old

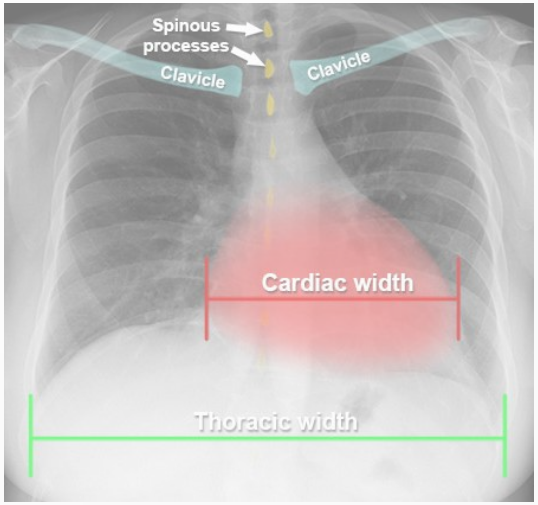


<1 years old

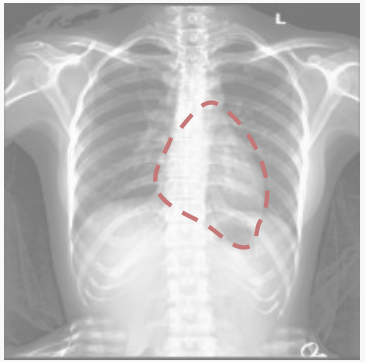
1. Infants have **larger hearts relative to their chest**
2. Cardiomegaly is **less common in children**

RQ: Would adult AI generalizes to children anyways?

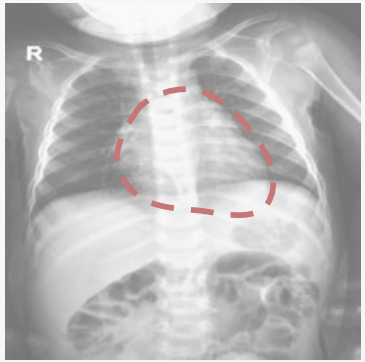
Case study. **Cardiomegaly** classification



enlarged heart



23 years old



<1 years old

1. Infants have **larger hearts relative to their chest**
2. Cardiomegaly is **less common in children**

Would a cardiomegaly classifier misclassify healthy children?

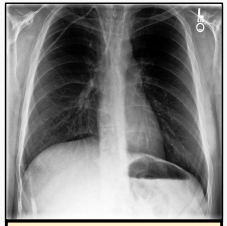
Case study. **Cardiomegaly**



VinDr-PCXR
Hanoi, Vietnam
***Ages**: 0-10 years old*



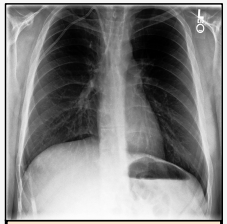
VinDr-CXR
Hanoi, Vietnam



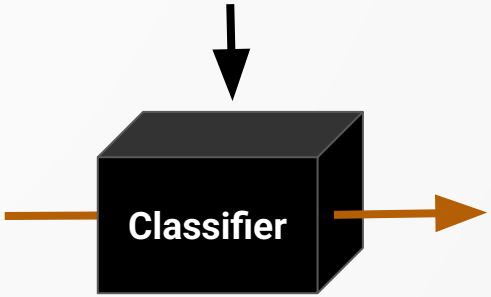
PadChest
Barcelona, Spain

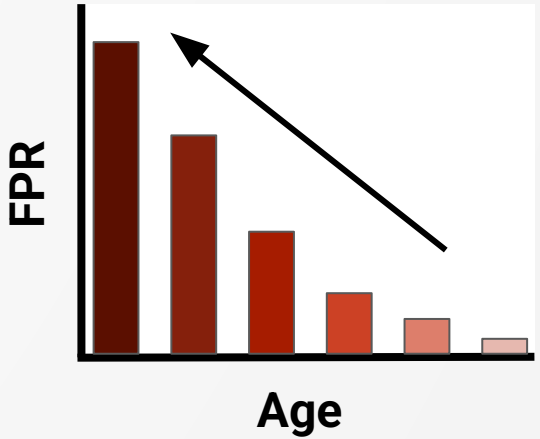


NIH X-ray14
USA (NIH)

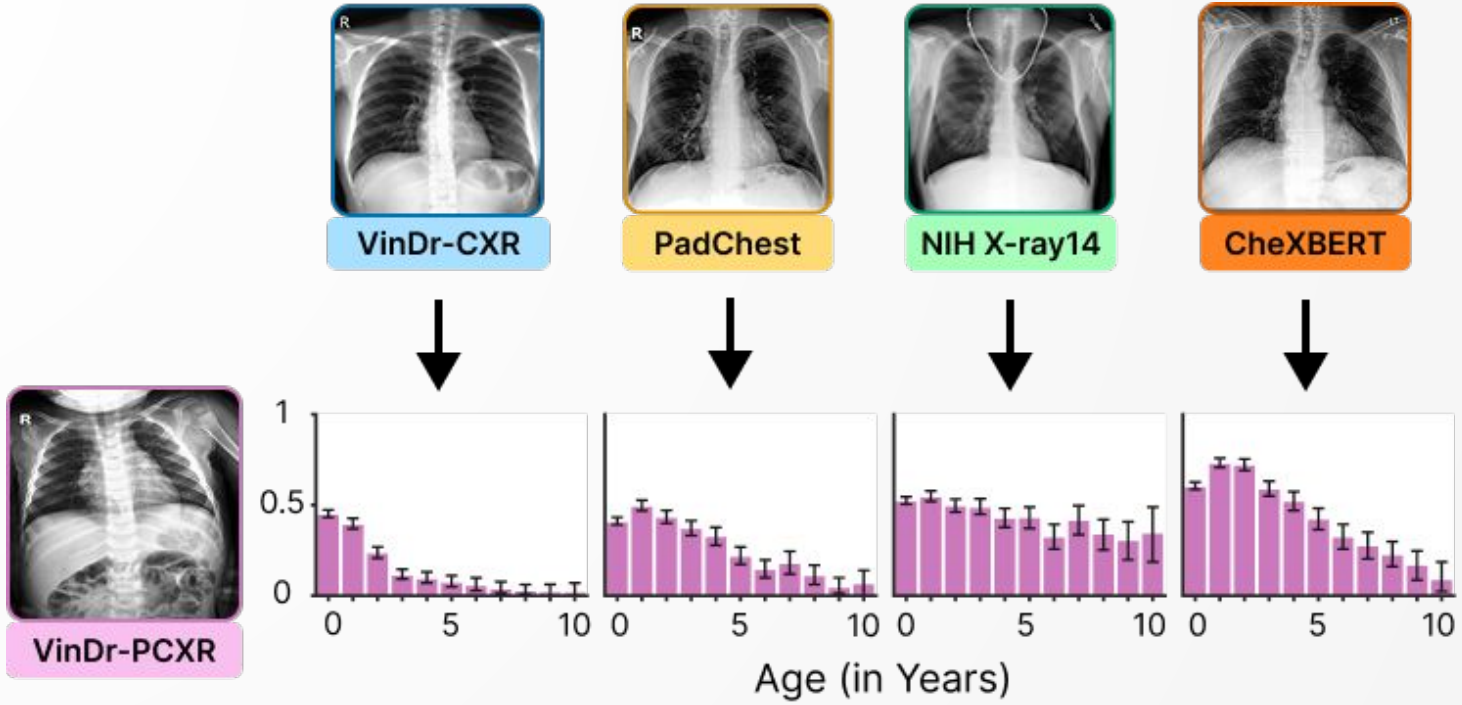


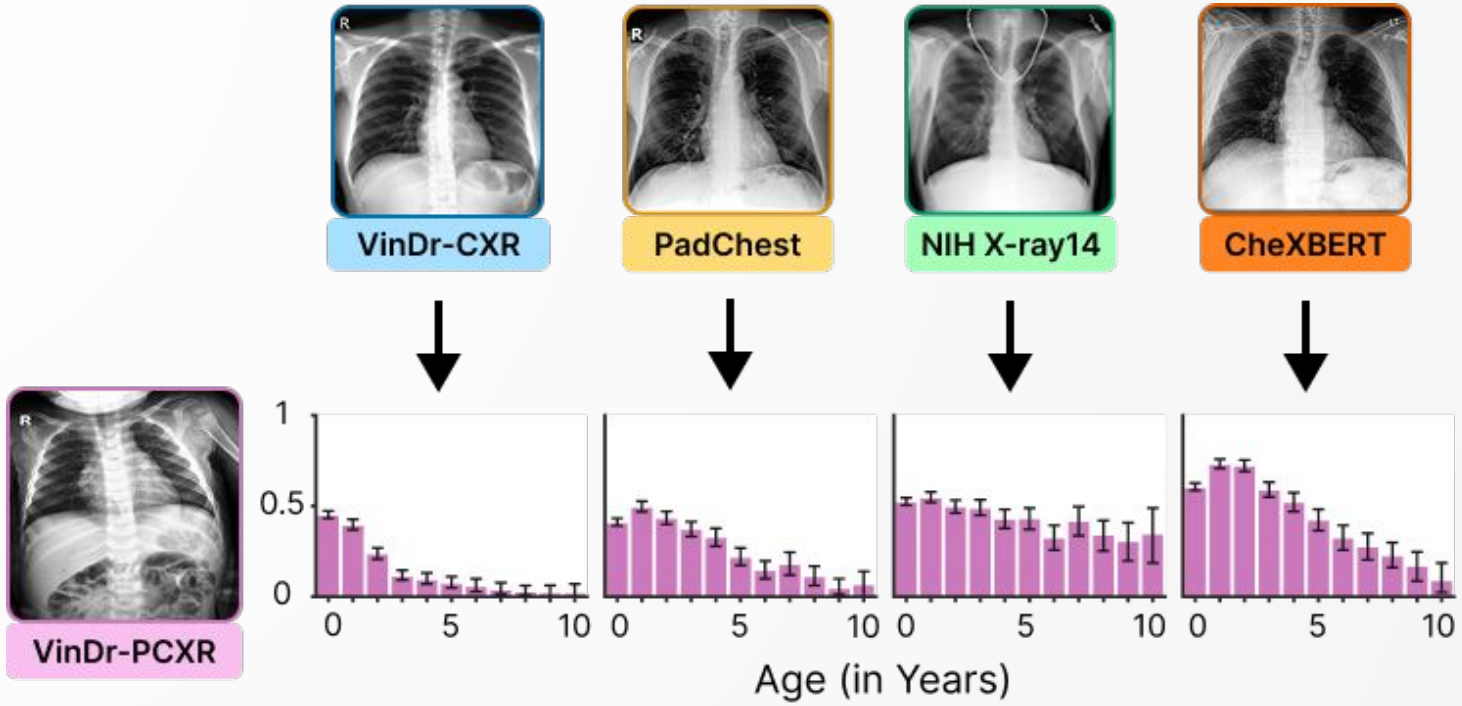
CheXBERT
USA (Stanford)





H_1: Adult AI fails on young healthy children


VinDr-CXR
PadChest
NIH X-ray14
CheXBERT
VinDr-PCXR
1
0.5
0
0
5
10
Age (in Years)



RQ: Would adult AI generalizes to children anyways?

No, and adult AI models may even **fail systematically**

Overview

1. children
2. public datasets
3. children in public datasets
4. **AI for ~~children~~ adults**

Where do we go from here?

Where do we go from here?

→ ~~Aim for at least __% children in public data~~

~~a) 0%~~ ~~b) 4%~~ ~~c) 30%~~ ~~d) 50%~~

no right answer.

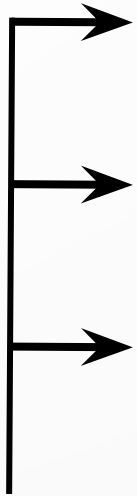
Where do we go from here?

→ Enough data for pediatric evaluation and re-training

→ Data to train AI for pediatric diseases

→ Patient age metadata in public datasets

Where do we go from here?

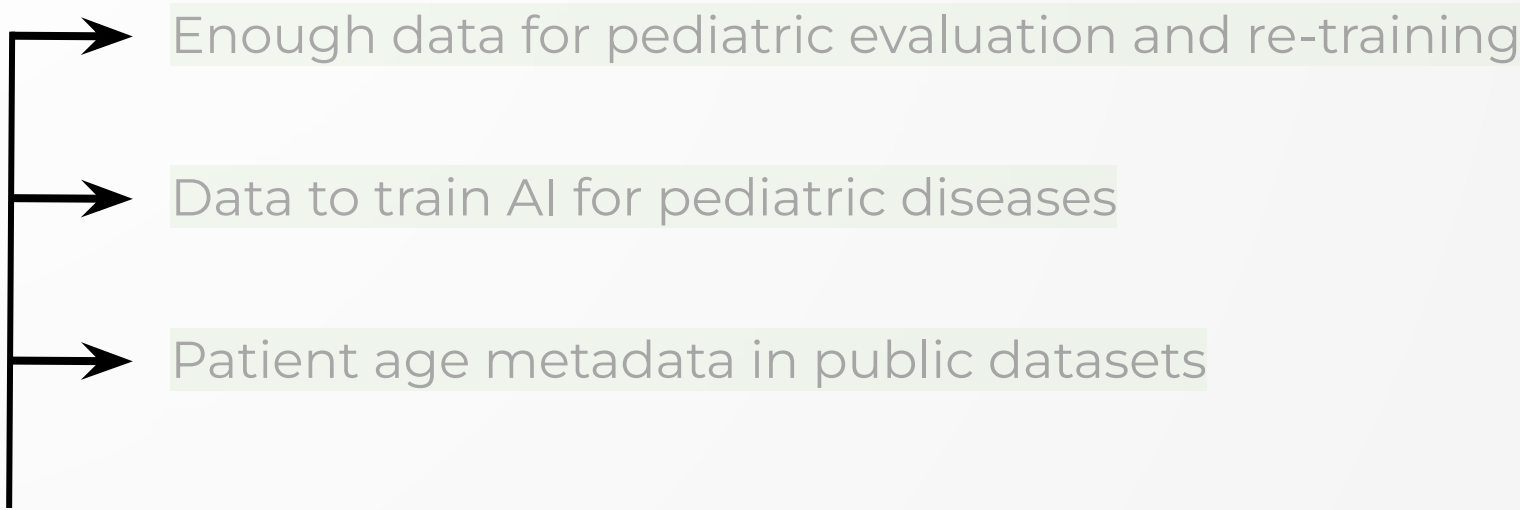


Enough data for pediatric evaluation and re-training

Data to train AI for pediatric diseases

Patient age metadata in public datasets

Where do we go from here?



1. Raise awareness on this issue → influence funders and policy-makers
2. Collaborate with pediatricians to develop AI for children

Acknowledgements



Alex Lu



Lauren Erdman



Alex Towbin



Irene Chen



Nicholas Heller

Questions?

Group Activity.

You’re now the head of the FDA, and your first task is to re-define the standard for an adult AI model to be approved for use on a child.

1. Would you require a randomized trial?
    a. If so, how would you recruit representative patients?
    b. If not, how would you deal with potential selection bias in observational data?
2. Would you define a blanket age cut-off? If not, what would you do?
3. Would you change the stringency based on the potential risk of the AI model?
    a. High risk: clinical diagnosis or treatment planning
    b. Low risk: administrative or clinical documentation
4. Would you involve a pediatrician? How?
5. How would you know in 2 years if what you did was wrong?