UCSF | UC Berkeley Computational Precision Health
SickKids Centre for Computational Medicine
Cleveland Clinic
Cincinnati Children's
Microsoft Research

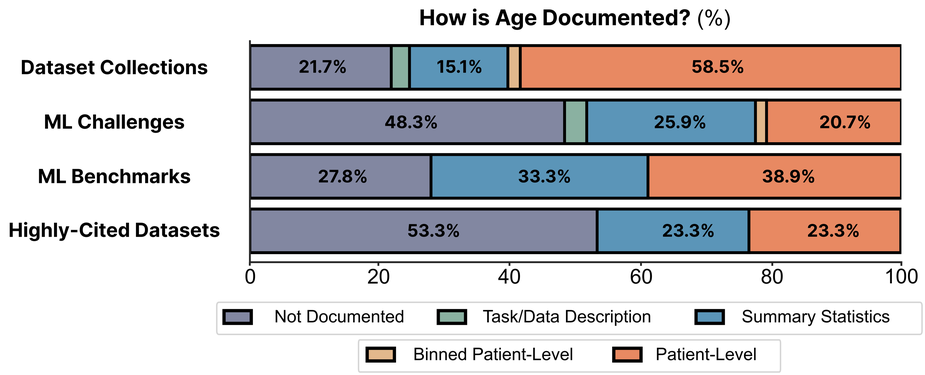
Are children present in the public medical imaging datasets supporting AI development?

Stanley Hua[1,2], Nicholas Heller[3], Alexander Towbin[4], Irene Chen[1], Alex Lu[5], Lauren Erdman[4]

There are fewer pediatric AI studies and FDA approvals for children!

Perhaps this disparity be partly explained by the landscape of public datasets? **Let's investigate pediatric representation in public medical imaging datasets and its implications.**

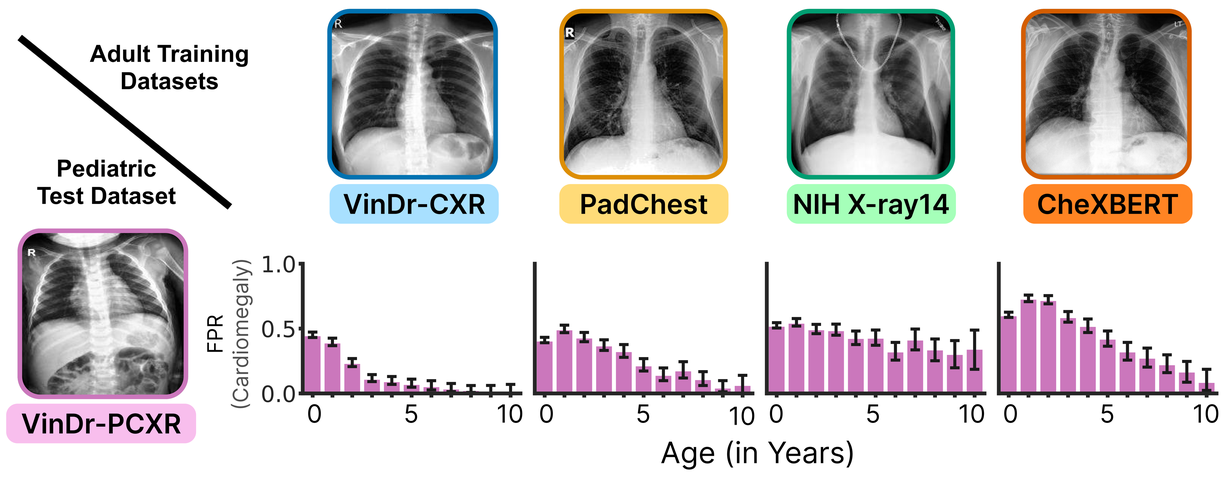
1 A third of public datasets fail to report patient ages, preventing age-stratified analysis.



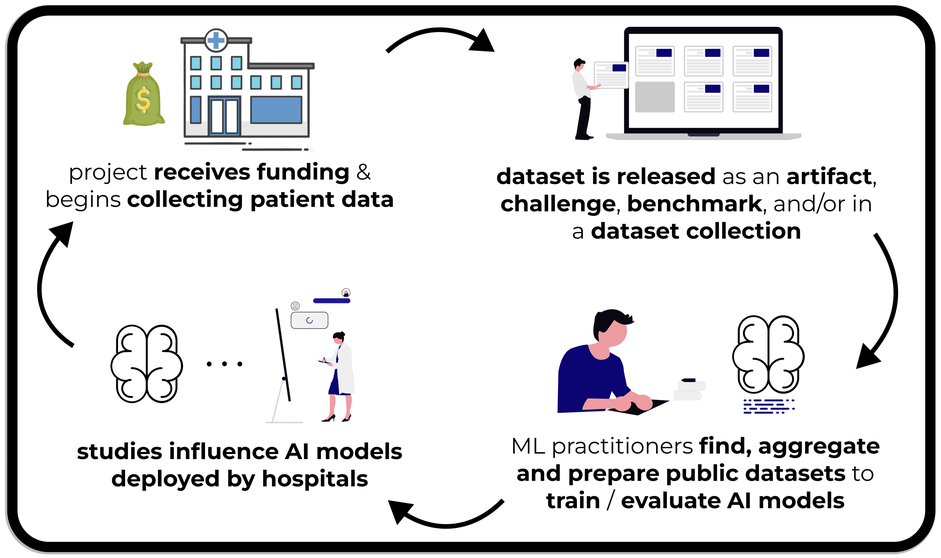
2 Children represent less than 2% of patients in public medical imaging datasets.

	Countries	Institutions	Datasets		Patients		
			All	Filtered	All	Children	%
Dataset Collections	11	65	106	80	231,077	4043	1.75
TCIA	11	61	69	51	42,145	478	1.14
Stanford AIMI	1	2	13	10	111,368	2096	1.88
MIDRC	1	3	24	19	77,565	1468	1.89
ML Challenges	28	100	58	29	101,181	212	0.21
Grand Challenge	16	53	44	24	74072	212	0.29
Kaggle	18	51	13	5	27110	0	0.0
Highly-Cited Datasets	14	40	30	19	235,478	1888	0.8
ML Benchmarks	14	37	18	15	142,721	83	0.06
MSD	11	29	10	7	2013	0	0.0
MedFAIR	4	9	7	7	140,108	76	0.05
AMOS	1	2	1	1	600	6	1.1
Total	34	181	203	135	512,608	6026	1.18
w/o Cancer	28	105	100	66	463,180	5883	1.27

3 Training AI models on adult-only public datasets can consistently lead to age bias.



4 To develop safe pediatric AI, we call on policy-makers, healthcare institutions and researchers to incentivize and prioritize the assembly and release of ML-ready pediatric datasets for training and evaluation.



Paper

Contact