KING'S College LONDON

FAIMI

Fairness of AI in Medical Imaging

RISE-MICCAI

Bias and Fairness in AI for Medical Imaging

Dr Andrew King

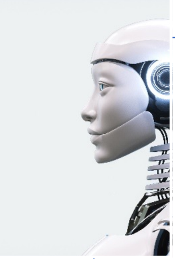
Biomedical Engineering Dept,

King's College London

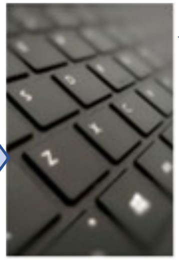
andrew.king@kcl.ac.uk

www.kclmmag.org

Talk overview



AI and the need for fairness



Definitions and terminology

Fair AI in medicine

A brief tour

What is (and isn't) Fair AI

Summary and resources

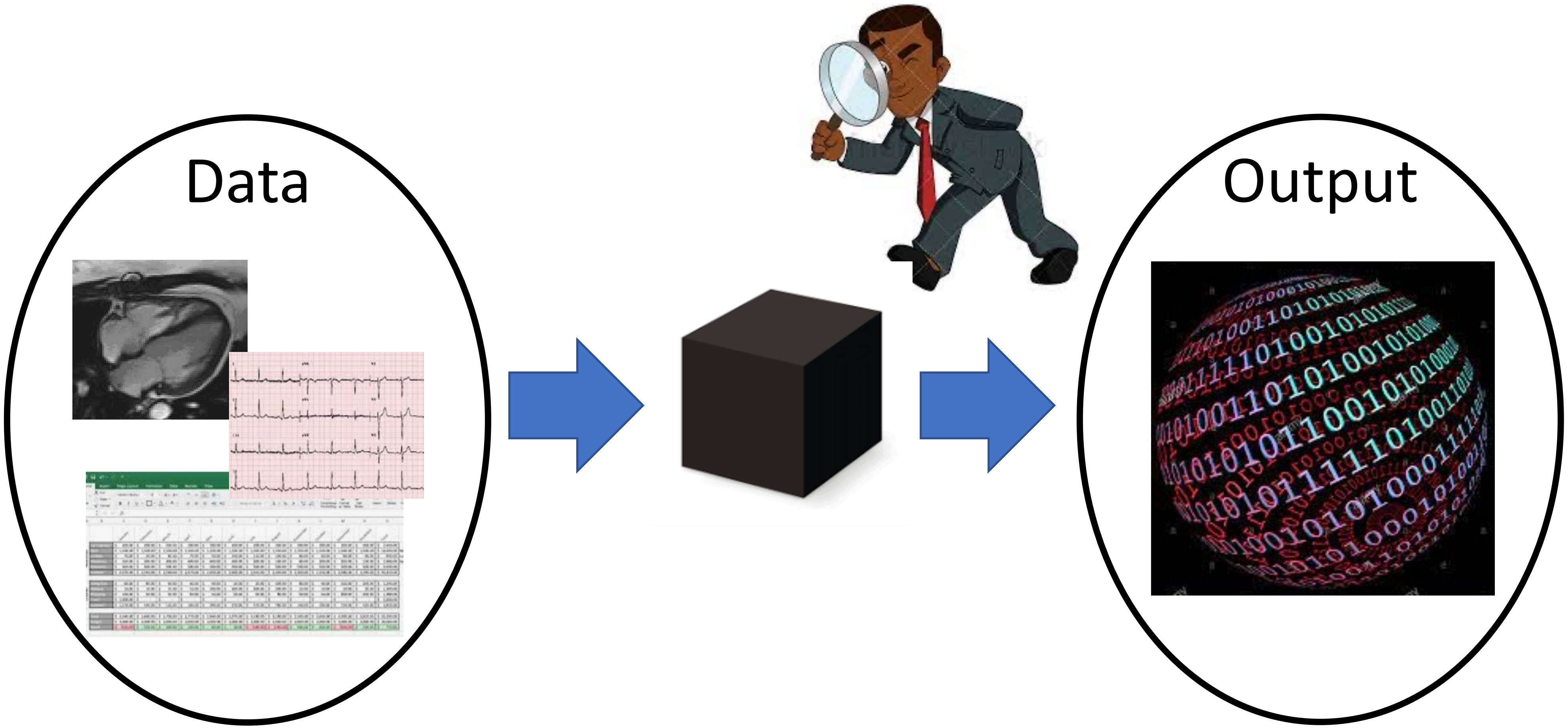
Learning objectives

- At the end of this session, you should be able to:
    - Explain the meaning of fairness/bias in AI, and describe why it has become an important issue recently
    - Explain some fairness definitions using common terminology for talking about bias in AI
    - Explain some approaches for addressing or mitigating bias in AI
    - Discuss some key papers on fairness in AI in medicine and beyond
    - Describe the differences between fair AI in medicine and other applications

AI and the need for fairness

AI models are "black boxes" …?

- What is a "black box" AI model?



→ *A model that cannot be (easily) understood/inspected by a human*

Are any AI models “black boxes”

In reality, in most cases we *can* inspect black box AI models …

0.070764	0.815829	0.621913	0.005636	0.38218	0.523031	0.377656	0.957161	0.565273	0.167925	0.746165	0.456912	0.77572	0.784241	0.540468	0.244342	0.231997	0.250309	0.134221	0.907605	0.592563	0.80369
0.483321	0.532398	0.394064	0.000413	0.715047	0.342433	0.367041	0.988568	0.328547	0.438413	0.928935	0.860191	0.334391	0.190375	0.505463	0.04037	0.570888	0.219539	0.291528	0.022548	0.17381	0.571508
0.025654	0.637817	0.745264	0.209306	0.794086	0.245243	0.74929	0.74318	0.086318	0.021706	0.658545	0.795351	0.584123	0.378427	0.16673	0.924719	0.450823	0.87991	0.488903	0.233867	0.464188	0.258668
0.458739	0.855065	0.812794	0.214675	0.213009	0.519728	0.251608	0.899683	0.021799	0.247867	0.484844	0.251849	0.640709	0.391501	0.087309	0.921581	0.658979	0.685022	0.10893	0.018061	0.824615	0.75709
0.686601	0.496198	0.734788	0.812598	0.279111	0.792759	0.788848	0.304802	0.348749	0.104745	0.968834	0.436956	0.610684	0.298782	0.880704	0.871662	0.248526	0.551364	0.071412	0.44075	0.925214	0.196381
0.163946	0.745473	0.632584	0.248179	0.549853	0.264346	0.254401	0.136947	0.655748	0.892907	0.099896	0.906641	0.344097	0.19098	0.557943	0.595705	0.363839	0.374378	0.316888	0.0658	0.199191	0.694628
0.265224	0.975943	0.615299	0.43226	0.598587	0.223069	0.19621	0.347113	0.542077	0.687871	0.177711	0.424957	0.777714	0.110775	0.35727	0.885367	0.890521	0.939376	0.19665	0.597822	0.139307	0.844147
0.828021	0.831173	0.78356	0.506504	0.715059	0.975438	0.66194	0.370643	0.541994	0.927797	0.680633	0.585556	0.858802	0.596126	0.869713	0.342543	0.900773	0.55927	0.116394	0.384397	0.434592	0.424681
0.442336	0.95272	0.752146	0.73247	0.196424	0.460062	0.668883	0.205808	0.927704	0.601753	0.903094	0.972625	0.390605	0.814197	0.132682	0.71869	0.745671	0.523139	0.634872	0.501049	0.858503	0.89899
0.800486	0.119112	0.232784	0.265336	0.009701	0.225551	0.344147	0.415407	0.50832	0.56056	0.187116	0.510336	0.990611	0.777314	0.46934	0.432835	0.866599	0.449885	0.713625	0.591951	0.183539	0.259877
0.248058	0.645941	0.079892	0.619505	0.715688	0.410693	0.980904	0.477146	0.835592	0.231674	0.172024	0.444635	0.069526	0.832643	0.248569	0.255224	0.81232	0.928535	0.574806	0.166288	0.942358	0.285104
0.887758	0.425867	0.745444	0.711243	0.20787	0.596284	0.591111	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	…9854	0.248375	0.602472	0.267257	0.364888	0.106645	0.218424	0.037235
0.372202	0.132537	0.643306	0.227898	0.642938	0.406558	0.740736	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	…5764	0.525438	0.428375	0.723777	0.103758	0.154232	0.444574	0.189415
0.337788	0.940276	0.065527	0.940405	0.274411	0.408014	0.266441	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	…5062	0.306428	0.018018	0.840317	0.108345	0.337278	0.129719	0.167052
0.871235	0.645879	0.206487	0.482292	0.673796	0.507118	0.439581	0.363178	0.462025	0.684408	0.720402	0.390506	0.908542	0.060076	0.16583	0.847212	0.265709	0.41061	0.110424	0.669831	0.82331	0.86471
0.41068	0.921559	0.971406	0.885903	0.035811	0.613762	0.614315	0.232069	0.734192	0.516054	0.70856	0.773233	0.222607	0.929548	0.907695	0.618914	0.620394	0.306957	0.29971	0.980206	0.334093	0.590254
0.364117	0.48634	0.353057	0.117043	0.196615	0.776124	0.797143	0.271933	0.861707	0.184235	0.034437	0.355614	0.241704	0.041842	0.906462	0.104335	0.841215	0.204549	0.561865	0.306735	0.402114	0.844216
0.428222	0.329656	0.929311	0.139111	0.716109	0.532672	0.259766	0.488198	0.763194	0.464763	0.610393	0.037875	0.339102	0.911369	0.545193	0.487161	0.99619	0.295575	0.888324	0.087538	0.227192	0.007543
0.468833	0.915261	0.380664	0.758698	0.276279	0.100982	0.635891	0.547674	0.67069	0.446553	0.741364	0.39475	0.250018	0.678715	0.927126	0.382334	0.840502	0.23363	0.343112	0.016357	0.934231	0.363493
0.811236	0.631483	0.9331	0.483219	0.473312	0.115413	0.618807	0.816561	0.840982	0.586519	0.227296	0.820013	0.147685	0.858761	0.526587	0.552791	0.793766	0.332453	0.028762	0.474096	0.438122	0.573262
0.588173	0.567702	0.535451	0.025501	0.971445	0.453778	0.984742	0.084767	0.115013	0.854615	0.420469	0.815032	0.625386	0.239217	0.479599	0.166119	0.551555	0.228466	0.082803	0.259113	0.822255	0.684583
0.185047	0.771555	0.968835	0.341976	0.898428	0.551813	0.493919	0.294589	0.5689	0.219123	0.464706	0.705371	0.870349	0.557707	0.097595	0.601709	0.009042	0.88915	0.585625	0.671803	0.63574	0.300568
0.826611	0.645304	0.820664	0.708642	0.562934	0.19499	0.355528	0.181336	0.505034	0.076119	0.400031	0.734987	0.150367	0.219003	0.631782	0.726244	0.380473	0.033212	0.882168	0.231485	0.412136	0.553205
0.577666	0.788134	0.414896	0.359781	0.003688	0.146451	0.847389	0.965508	0.810842	0.537562	0.25671	0.958786	0.251096	0.250755	0.439882	0.994636	0.318212	0.063905	0.47202	0.2195	0.98684	0.810641
0.167326	0.49	0.362958	0.624156	0.03584	0.091384	0.356056	0.686149	0.370071	0.29601	0.710059	0.067027	0.208281	0.599449	0.128647	0.252855	0.76348	0.379109	0.782016	0.104975	0.230784	0.943101
0.203525	0.29204	0.382315	0.848286	0.572992	0.69466	0.921583	0.055079	0.359713	0.32363	0.146071	0.846933	0.583479	0.530763	0.154313	0.290768	0.267591	0.857319	0.475032	0.878925	0.745705	0.079983
0.014113	0.188175	0.115506	0.34473	0.02383	0.8905	0.541874	0.964236	0.019362	0.839808	0.064341	0.65785	0.113326	0.158616	0.749311	0.657631	0.112312	0.03724	0.896456	0.414804	0.97036	0.579378
0.688415	0.695073	0.532504	0.203141	0.180844	0.228554	0.695813	0.488019	0.757916	0.332769	0.01145	0.973697	0.966756	0.543618	0.12745	0.629102	0.040751	0.575509	0.679101	0.850549	0.000827	0.515741

But can we *understand* them?

Are *all* AI models "black boxes"?

- "White box" AI: refers to models that can be *understood* by humans

0.070764	0.815829	0.621913	0.005636	0.38218	0.523031	0.377656	0.957161	0.565273	0.167925	0.746165	0.456912	0.77572	0.784241	0.540468	0.244342	0.231997	0.250309	0.134221	0.907605	0.592563	0.80369
0.483321	0.532398	0.394064	0.000413	0.715047	0.342433	0.367041	0.988568	0.328547	0.438413	0.928935	0.860191	0.334391	0.190375	0.505463	0.04037	0.570888	0.219539	0.291528	0.022548	0.17381	0.571508
0.025654	0.637817	0.745264	0.209306	0.794086	0.245243	0.74929	0.74318	0.086318	0.021706	0.658545	0.795351	0.584123	0.378427	0.16673	0.924719	0.450823	0.87991	0.488903	0.233867	0.464188	0.258668
0.458739	0.855065	0.812794	0.214675	0.213009	0.519728	0.251608	0.899683	0.021799	0.247867	0.484844	0.251849	0.640709	0.391501	0.087309	0.921581	0.658979	0.685022	0.10893	0.018061	0.824615	0.75709
0.686601	0.496198	0.734788	0.812598	0.279111	0.792759	0.788848	0.304802	0.348749	0.104745	0.968834	0.436956	0.610684	0.298782	0.880704	0.871662	0.248526	0.551364	0.071412	0.44075	0.925214	0.196381
0.163946	0.745473	0.632584	0.248179	0.549853	0.264346	0.254401	0.136947	0.655748	0.892907	0.099896	0.906641	0.344097	0.19098	0.557943	0.595705	0.363839	0.374378	0.316888	0.0658	0.199191	0.694628
0.265224	0.975943	0.615299	0.43226	0.598587	0.223069	0.19621	0.347113	0.542077	0.687871	0.177711	0.424957	0.777714	0.110775	0.35727	0.885367	0.890521	0.939376	0.19665	0.597822	0.139307	0.844147
0.828021	0.831173	0.78356	0.506504	0.715059	0.975438	0.66194	0.370643	0.541994	0.927797	0.680633	0.585556	0.858802	0.596126	0.869713	0.342543	0.900773	0.55927	0.116394	0.384397	0.434592	0.424681
0.442336	0.95272	0.752146	0.73247	0.196424	0.460062	0.668883	0.205808	0.927704	0.601753	0.903094	0.972625	0.390605	0.814197	0.132682	0.71869	0.745671	0.523139	0.634872	0.501049	0.858503	0.89899
0.800486	0.119112	0.232784	0.265336	0.009701	0.225551	0.344147	0.415407	0.50832	0.56056	0.187116	0.510336	0.990611	0.777314	0.46934	0.432835	0.866599	0.449885	0.713625	0.591951	0.183539	0.259877
0.248058	0.645941	0.079892	0.619505	0.715688	0.410693	0.980904	0.477146	0.835592	0.231674	0.172024	0.444635	0.069526	0.832643	0.248569	0.255224	0.81232	0.928535	0.574806	0.166288	0.942358	0.285104
0.887758	0.425867	0.745444	0.711243	0.20787	0.596284	0.591111	0.504856	0.679362	0.10328	0.06886	0.995714	0.669428	0.862282	0.349854	0.248375	0.602472	0.267257	0.364888	0.106645	0.218424	0.037235
0.372202	0.132537	0.643306	0.227898	0.642938	0.406558	0.740736	0.554679	0.786813	0.69206	0.089457	0.133662	0.223231	0.373996	0.365764	0.525438	0.428375	0.723777	0.103758	0.154232	0.444574	0.189415
0.337788	0.940276	0.065527	0.940405	0.274411	0.408014	0.266441	0.823756	0.389341	0.097469	0.960532	0.5834	0.435949	0.328611	0.86062	0.306428	0.018018	0.840317	0.108345	0.337278	0.129719	0.167052
0.871235	0.645879	0.206487	0.482292	0.673796	0.507118	0.439581	0.363178	0.462025	0.684408	0.720402	0.390506	0.908542	0.060076	0.16583	0.847212	0.265709	0.41061	0.110424	0.669831	0.82331	0.86471
0.41068	0.921559	0.971406	0.885903	0.035811	0.613762	0.614315	0.232069	0.734192	0.516054	0.70856	0.773233	0.222607	0.929548	0.907695	0.618914	0.620394	0.306957	0.29971	0.980206	0.334093	0.590254
0.364117	0.48634	0.353057	0.117043	0.196615	0.776124	0.797143	0.271933	0.861707	0.184235	0.034437	0.355614	0.241704	0.041842	0.906462	0.104335	0.841215	0.204549	0.561865	0.306735	0.402114	0.844216
0.428222	0.329656	0.929311	0.139111	0.716109	0.532672	0.259766	0.488198	0.763194	0.464763	0.610393	0.037875	0.339102	0.911369	0.545193	0.487161	0.99619	0.295575	0.888324	0.087538	0.227192	0.007543
0.468833	0.915261	0.380664	0.758698	0.276279	0.100982	0.635891	0.547674	0.67069	0.446553	0.741364	0.39475	0.250018	0.678715	0.927126	0.382334	0.840502	0.23363	0.343112	0.016357	0.934231	0.363493
0.811236	0.631483	0.9331	0.483219	0.473312	0.115413	0.618807	0.816561	0.840982	0.586519	0.227296	0.820013	0.147685	0.858761	0.526587	0.552791	0.793766	0.332453	0.028762	0.474096	0.438122	0.573262
0.588173	0.567702	0.535451	0.025501	0.971445	0.453778	0.984742	0.084767	0.115013	0.854615	0.420469	0.815032	0.625386	0.239217	0.479599	0.166119	0.551555	0.228466	0.082803	0.259113	0.822255	0.684583
0.185047	0.771555	0.968835	0.341976	0.898428	0.551813	0.493919	0.294589	0.5689	0.219123	0.464706	0.705371	0.870349	0.557707	0.097595	0.601709	0.009042	0.88915	0.585625	0.671803	0.63574	0.300568
0.826611	0.645304	0.820664	0.708642	0.562934	0.19499	0.355528	0.181336	0.505034	0.076119	0.400031	0.734987	0.150367	0.219003	0.631782	0.726244	0.380473	0.033212	0.882168	0.231485	0.412136	0.553205
0.577666	0.788134	0.414896	0.359781	0.003688	0.146451	0.847389	0.965508	0.810842	0.537562	0.25671	0.958786	0.251096	0.250755	0.439882	0.994636	0.318212	0.063905	0.47202	0.2195	0.98684	0.810641
0.167326	0.49	0.362958	0.624156	0.03584	0.091384	0.356056	0.686149	0.370071	0.29601	0.710059	0.067027	0.208281	0.599449	0.128647	0.252855	0.76348	0.379109	0.782016	0.104975	0.230784	0.943101
0.203525	0.29204	0.382315	0.848286	0.572992	0.69466	0.921583	0.055079	0.359713	0.32363	0.146071	0.846933	0.583479	0.530763	0.154313	0.290768	0.267591	0.857319	0.475032	0.878925	0.745705	0.079983
0.014113	0.188175	0.115506	0.34473	0.02383	0.8905	0.541874	0.964236	0.019362	0.839808	0.064341	0.65785	0.113326	0.158616	0.749311	0.657631	0.112312	0.03724	0.896456	0.414804	0.97036	0.579378
0.688415	0.695073	0.532504	0.203141	0.180844	0.228554	0.695813	0.488019	0.757916	0.332769	0.01145	0.973697	0.966756	0.543618	0.12745	0.629102	0.040751	0.575509	0.679101	0.850549	0.000827	0.515741

Medium priority **High priority**

White box AI

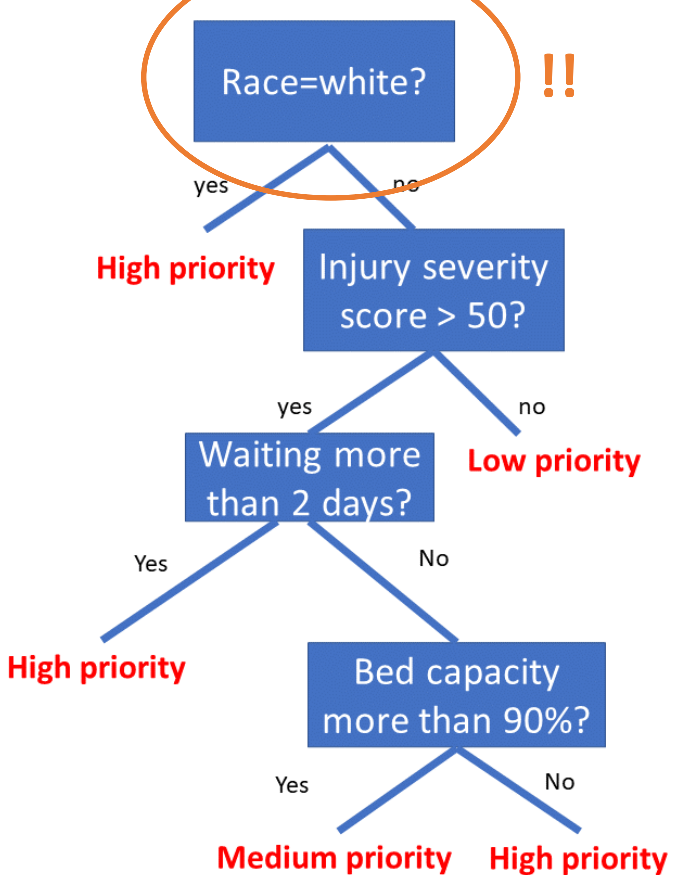
- *White box models* aim to make AI models more “transparent”, i.e. to enable humans to understand better how outputs were produced
    - We can also call this *interpretability* or *explainability*

- Why is *white box AI* useful?

Why can white box AI be useful?

White box model

Prioritising Urgent Care Patients for Bed Allocation



Black box model

0.070764	0.815829	0.621913	0.005636	0.38218	0.523031	0.377656	0.957161	0.565273	0.167925	0.746165	0.456912	0.77572	0.784241	0.540468	0.244342	0.231997	0.250309	0.134221	0.907605	0.592563	0.80369
0.483321	0.532398	0.394064	0.000413	0.715047	0.342433	0.367041	0.988568	0.328547	0.438413	0.928935	0.860191	0.334391	0.190375	0.505463	0.04037	0.570888	0.219539	0.291528	0.022548	0.17381	0.571508
0.025654	0.637817	0.745264	0.209306	0.794086	0.245243	0.74929	0.74318	0.086318	0.021706	0.658545	0.795351	0.584123	0.378427	0.16673	0.924719	0.450823	0.87991	0.488903	0.233867	0.464188	0.258668
0.458739	0.855065	0.812794	0.214675	0.213009	0.519728	0.251608	0.899683	0.021799	0.247867	0.484844	0.251849	0.640709	0.391501	0.087309	0.921581	0.658979	0.685022	0.10893	0.018061	0.824615	0.75709
0.686601	0.496198	0.734788	0.812598	0.279111	0.792759	0.788848	0.304802	0.348749	0.104745	0.968834	0.436956	0.610684	0.298782	0.880704	0.871662	0.248526	0.551364	0.071412	0.44075	0.925214	0.196381
0.163946	0.745473	0.632584	0.248179	0.549853	0.264346	0.254401	0.136947	0.655748	0.892907	0.099896	0.906641	0.344097	0.19098	0.557943	0.595705	0.363839	0.374378	0.316888	0.0658	0.199191	0.694628
0.265224	0.975943	0.615299	0.43226	0.598587	0.223069	0.19621	0.347113	0.542077	0.687871	0.177711	0.424957	0.777714	0.110775	0.35727	0.885367	0.890521	0.939376	0.19665	0.597822	0.139307	0.844147
0.828021	0.831173	0.78356	0.506504	0.715059	0.975438	0.66194	0.370643	0.541994	0.927797	0.680633	0.585556	0.858802	0.596126	0.869713	0.342543	0.900773	0.55927	0.116394	0.384397	0.434592	0.424681
0.442336	0.95272	0.752146	0.73247	0.196424	0.460062	0.668883	0.205808	0.927704	0.601753	0.903094	0.972625	0.390605	0.814197	0.132682	0.71869	0.745671	0.523139	0.634872	0.501049	0.858503	0.89899
0.800486	0.119112	0.232784	0.265336	0.009701	0.225551	0.344147	0.415407	0.50832	0.56056	0.187116	0.510336	0.990611	0.777314	0.46934	0.432835	0.866599	0.449885	0.713625	0.591951	0.183539	0.259877
0.248058	0.645941	0.079892	0.619505	0.715688	0.410693	0.980904	0.477146	0.835592	0.231674	0.172024	0.444635	0.069526	0.832643	0.248569	0.255224	0.81232	0.928535	0.574806	0.166288	0.942358	0.285104
0.887758	0.425867	0.745444	0.711243	0.20787	0.596284	0.591111	0.504856	0.679362	0.10328	0.06886	0.995714	0.669428	0.862282	0.349854	0.248375	0.602472	0.267257	0.364888	0.106645	0.218424	0.037235
0.372202	0.132537	0.643306	0.227898	0.642938	0.406558	0.740736	0.554679	0.786813	0.69206	0.089457	0.133662	0.223231	0.373996	0.365764	0.525438	0.428375	0.723777	0.103758	0.154232	0.444574	0.189415
0.337788	0.940276	0.065527	0.940405	0.274411	0.408014	0.266441	0.823756	0.389341	0.097469	0.960532	0.5834	0.435949	0.328611	0.86062	0.306428	0.018018	0.840317	0.108345	0.337278	0.129719	0.167052
0.871235	0.645879	0.206487	0.482292	0.673796	0.507118	0.439581	0.363178	0.462025	0.684408	0.720402	0.390506	0.908542	0.060076	0.16583	0.847212	0.265709	0.41061	0.110424	0.669831	0.82331	0.86471
0.41068	0.921559	0.971406	0.885903	0.035811	0.613762	0.614315	0.232069	0.734192	0.516054	0.70856	0.773233	0.222607	0.929548	0.907695	0.618914	0.620394	0.306957	0.29971	0.980206	0.334093	0.590254
0.364117	0.48634	0.353057	0.117043	0.196615	0.776124	0.797143	0.271933	0.861707	0.184235	0.034437	0.355614	0.241704	0.041842	0.906462	0.104335	0.841215	0.204549	0.561865	0.306735	0.402114	0.844216
0.428222	0.329656	0.929311	0.139111	0.716109	0.532672	0.259766	0.488198	0.763194	0.464763	0.610393	0.037875	0.339102	0.911369	0.545193	0.487161	0.99619	0.295575	0.888324	0.087538	0.227192	0.007543
0.468833	0.915261	0.380664	0.758698	0.276279	0.100982	0.635891	0.547674	0.67069	0.446553	0.741364	0.39475	0.250018	0.678715	0.927126	0.382334	0.840502	0.23363	0.343112	0.016357	0.934231	0.363493
0.811236	0.631483	0.9331	0.483219	0.473312	0.115413	0.618807	0.816561	0.840982	0.586519	0.227296	0.820013	0.147685	0.858761	0.526587	0.552791	0.793766	0.332453	0.028762	0.474096	0.438122	0.573262
0.588173	0.567702	0.535451	0.025501	0.971445	0.453778	0.984742	0.084767	0.115013	0.854615	0.420469	0.815032	0.625386	0.239217	0.479599	0.166119	0.551555	0.228466	0.082803	0.259113	0.822255	0.684583
0.185047	0.771555	0.968835	0.341976	0.898428	0.551813	0.493919	0.294589	0.5689	0.219123	0.464706	0.705371	0.870349	0.557707	0.097595	0.601709	0.009042	0.88915	0.585625	0.671803	0.63574	0.300568
0.826611	0.645304	0.820664	0.708642	0.562934	0.19499	0.355528	0.181336	0.505034	0.076119	0.400031	0.734987	0.150367	0.219003	0.631782	0.726244	0.380473	0.033212	0.882168	0.231485	0.412136	0.553205
0.577666	0.788134	0.414896	0.359781	0.003688	0.146451	0.847389	0.965508	0.810842	0.537562	0.25671	0.958786	0.251096	0.250755	0.439882	0.994636	0.318212	0.063905	0.47202	0.2195	0.98684	0.810641
0.167326	0.49	0.362958	0.624156	0.03584	0.091384	0.356056	0.686149	0.370071	0.29601	0.710059	0.067027	0.208281	0.599449	0.128647	0.252855	0.76348	0.379109	0.782016	0.104975	0.230784	0.943101
0.203525	0.29204	0.382315	0.848286	0.572992	0.69466	0.921583	0.055079	0.359713	0.32363	0.146071	0.846933	0.583479	0.530763	0.154313	0.290768	0.267591	0.857319	0.475032	0.878925	0.745705	0.079983
0.014113	0.188175	0.115506	0.34473	0.02383	0.8905	0.541874	0.964236	0.019362	0.839808	0.064341	0.65785	0.113326	0.158616	0.749311	0.657631	0.112312	0.03724	0.896456	0.414804	0.97036	0.579378
0.688415	0.695073	0.532504	0.203141	0.180844	0.228554	0.695813	0.488019	0.757916	0.332769	0.01145	0.973697	0.966756	0.543618	0.12745	0.629102	0.040751	0.575509	0.679101	0.850549	0.000827	0.515741

White box models enable us to more easily uncover potential *bias* or *toxic behaviour*, and also address/remove it.

So do we have to use simple white box models?

- ***Problem***: the best-performing techniques, especially on more complex tasks, are *black box AI* models
- So how can we know if they are biased, and if they are what can we do about it?

$\longrightarrow$ $\boxed{\text{Fair AI …}}$

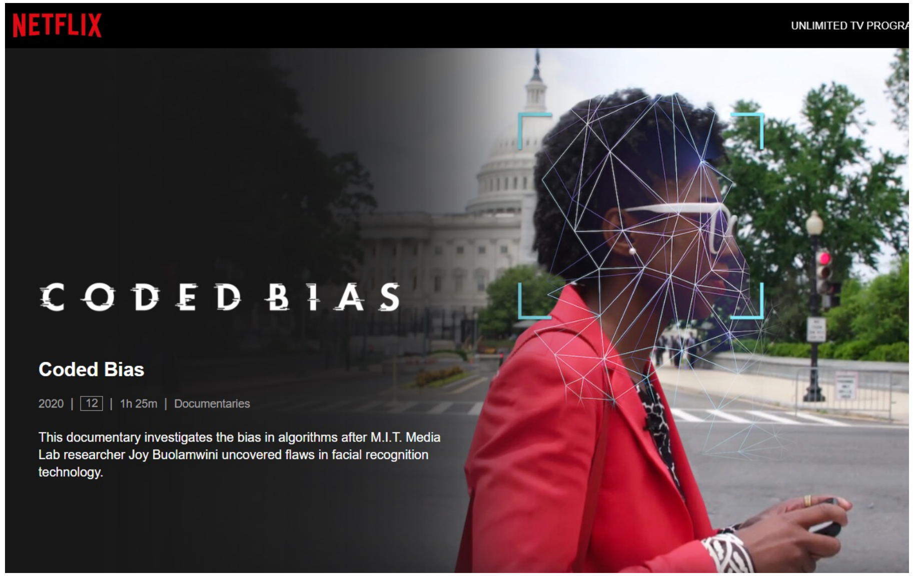
- Since the advent of black box models like *deep learning* we have seen two things:
  1. A significant *improvement* in *performance*
  2. A significant *reduction* in *interpretability*
- *Fair AI* is mostly a response to the second point (but also partly the first …)

The birth of Fair AI? Gender Shades …

- Landmark paper in 2018:

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139
Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011



Algorithmic Justice League:
www.ajl.org/

Distributed AI Research Institute:
www.dair-institute.org/

Home | Coronavirus | UK | World | Business | Politics | Tech | Science | Health | Family & Education | Entertai

Technology

Timnit Gebru: Google staff rally behind fired AI researcher

4 December

Hundreds of Google staff have signed a letter backing a leading AI ethics researcher who was sacked by Google.

Bias in image generation

Open AI's DALL-E 2 pictures of lawyers …

… and flight attendants

https://www.vox.com/future-perfect/23023538/ai-dalle-2-openai-bias-gpt-3-incentives

Definitions & terminology

Definitions and terminology

- In the context of decision-making, *fairness* refers to:
  - The "*absence of any prejudice or favouritism toward an individual or a group based on their inherent or acquired characteristics*"

- Fairness in AI refers to methods for *assessing* and/or *addressing* such prejudice/bias, i.e. ***bias = different levels of performance for different groups of subjects (e.g. races or sexes)***

<u>Terminology</u>:
Protected attribute: the characteristic(s) for which fairness needs to be ensured, e.g. sex, race (a.k.a. **sensitive attribute**)
Protected group: a set of samples with the same value(s) for the protected attribute(s)

Definitions of fairness

- Fairness can be defined in a number of different ways
- We will consider an AI classifier for detecting a disease (e.g. breast cancer) from a medical image
- Our protected groups are White subjects and Black subjects
- In the definitions, we will refer to the numbers of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN)

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Definitions of fairness – equalised odds

- Equalised odds:
  - *"… the probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for the protected … groups"*
  - I.e. equal true positive rate (TPR) & false positive rate (FPR) for each protected group
    - TPR = TP / (TP + FN) *(this is also known as the sensitivity)*
    - FPR = FP / (FP + TN) *(this is one minus the specificity)*

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

N. Mehrabi, et al. (2021) "A Survey on Bias and Fairness in Machine Learning", ACM Computing Surveys.

Definitions of fairness – demographic parity

- Demographic parity:
  - *"The likelihood of a positive outcome should be the same regardless of … the protected group"*
  - I.e. an equal chance of being classified positive for each protected group
    - Predicted positive rate =

      TN + FP / (TN + FP + TN + FN)

      (Note this does not depend on the ground truth)

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

N. Mehrabi, et al. (2021) "A Survey on Bias and Fairness in Machine Learning", ACM Computing Surveys.

Definitions of fairness – equal opportunity

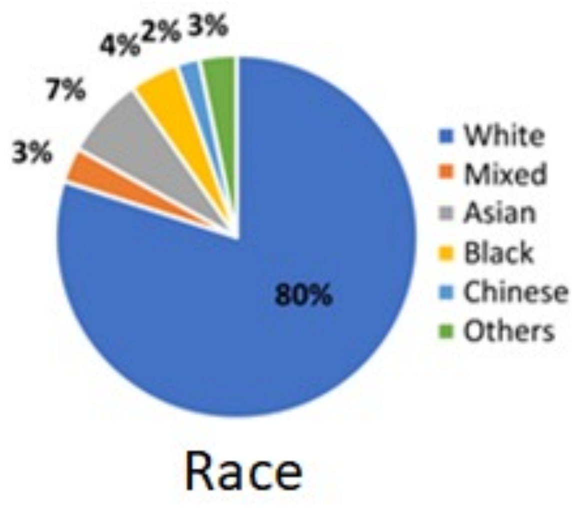
- Equal opportunity:
  - *"The probability of a person in a positive class being assigned to a positive outcome should be equal for both protected … group's members"*
  - I.e. the true positive rate (=sensitivity) should be the same for each protected group
    - TPR = TP / (TP + FN)
      (this is also known as the sensitivity)

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

N. Mehrabi, et al. (2021) "A Survey on Bias and Fairness in Machine Learning", ACM Computing Surveys.

Source of bias

- Bias is often (but not always) caused by an *imbalance* between *protected groups* in the *training data*
- E.g. this is the race breakdown in the widely-used UK Biobank dataset:



- When this level of imbalance is combined with *differences* in the data for different protected groups (e.g. races), it can cause the AI model to optimize its performance for the over-represented group(s)

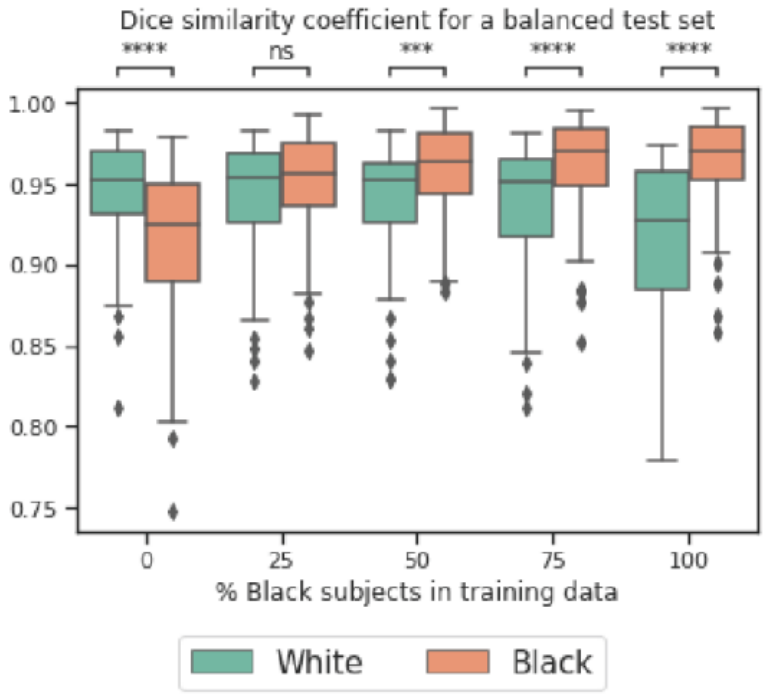
Addressing bias

Prepare data → Train model → Deploy model

Pre-Processing Algorithms Mitigate bias in training data	**In-Processing Algorithms** Mitigate bias in classifiers	**Post-Processing Algorithms** Mitigate bias in predictions
Reweighing Modifies the weights of different training examples	**Adversarial Debiasing** Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions	**Reject Option Classification** Changes predictions from a classifier to make them more fair
Disparate Impact Remover Edits feature values to improve group fairness	**Prejudice Remover** Adds a discrimination-aware regularization term to the learning objective	**Calibrated Equalized Odds** Optimizes over calibrated classifier score outputs that lead to fair output labels
Optimized Preprocessing Modifies training data features and labels	**Meta Fair Classifier** Takes the fairness metric as part of the input and returns a classifier optimized for the metric	**Equalized Odds** Modifies the predicted label using an optimization scheme to make predictions more fair
Learning Fair Representations Learns fair representations by obfuscating information about protected attributes		

Bias mitigation - preprocessing

- *Address data imbalance*:
  - If we can, add more training data for protected groups that are under-performing
  - If not, can use data augmentation
- *Reweighing*:
  - When computing the loss of an AI model, add a weighting to increase the importance of under-represented protected groups
- *Over-sampling*:
  - When sampling training data in minibatches, over-sample from the under-performing group



T. Lee, et al. (2022) "A Systematic Study of Race and Sex Bias in CNN-based Cardiac MR Segmentation", Proceedings MICCAI STACOM.

Approach	Segmentation					
	White	Mixed	Asian	Black	Chinese	Others
Baseline - Fairness through unawareness	93.51	84.52	88.90	85.88	87.63	85.66
1. Stratified batch sampling	90.88	93.84	93.65	93.07	94.35	93.50

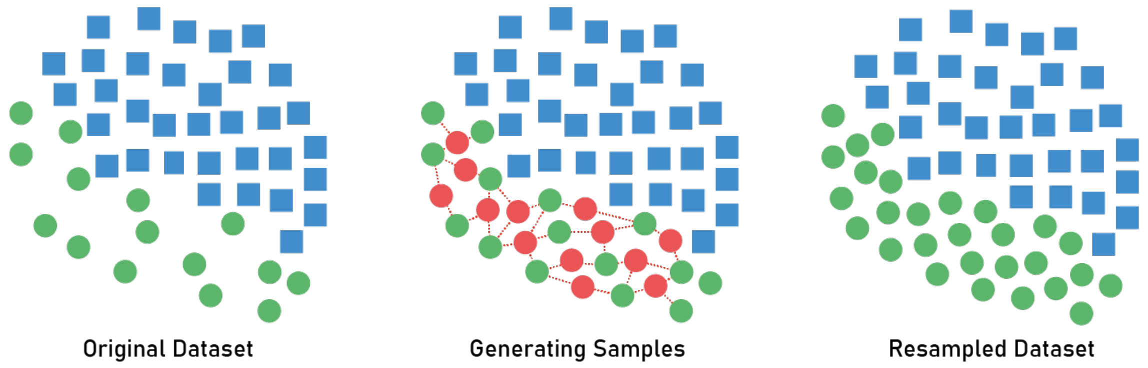
Puyol-Antón, et al, "Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation", Proceedings MICCAI, 2021

Bias mitigation – preprocessing: example

- *SMOTE: Synthetic Minority Oversampling Technique*
  - SMOTE is a widely used data augmentation technique in machine learning for imbalanced datasets
  - It works by creating synthetic samples of the minority class by interpolating between existing minority class instances

Steps in SMOTE

- **Step 1**: Select a random sample from the minority class
- **Step 2**: Choose one of its k-nearest neighbours
- **Step 3**: Create a synthetic instance by interpolating between the original sample and the neighbour
- **Step 4**: Repeat until the desired level of oversampling is achieved



- Normally used to address class imbalance problem, but can also be applied to boost representation of minority protected groups -> *Fair-SMOTE*

J. Chakraborty, et al. (2021) "Bias in Machine Learning Software: Why? How? What to Do?", Proceedings ESEC/FSE.

Bias mitigation - inprocessing

- *Meta-fair classifier*:
  - Co-train model for intended task with a protected attribute classifier model

Approach	Segmentation					
	White	Mixed	Asian	Black	Chinese	Others
Baseline - Fairness through unawareness	93.51	84.52	88.90	85.88	87.63	85.66
1. Stratified batch sampling	90.88	93.84	93.65	93.07	94.35	93.50
2. Fair meta-learning for segmentation	92.75	88.03	90.64	89.60	88.18	88.27

Puyol-Antón, et al, "Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation", Proceedings MICCAI, 2021

- *Representation neutralisation*:
  - *Modify loss to avoid encoding of protected attributes (e.g. adversarial training)*

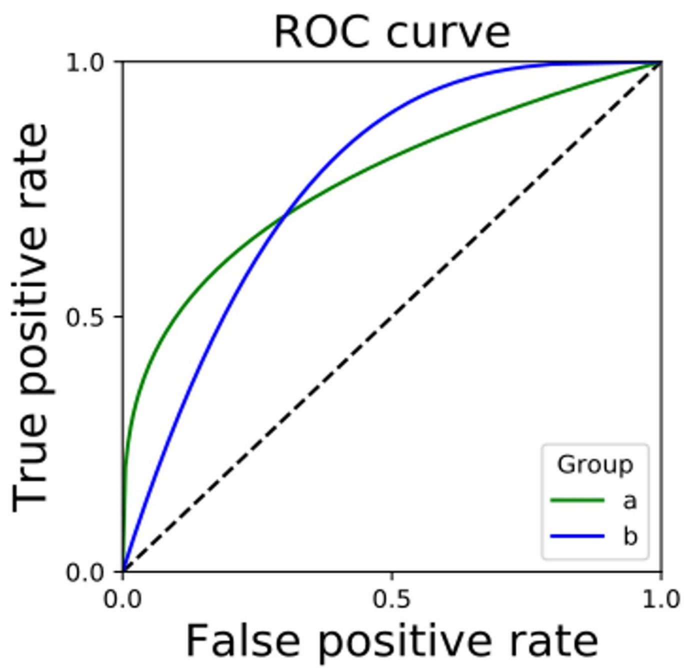
Du, et al, "Fairness via Representation Neutralization", Proceedings NeurIPS, 2021

- *Training for fairness*:
  - *Modify loss encourage a particular fairness criterion (e.g. Group DRO)*

Sagawa, et al, "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-case Generalization", Proceedings ICLR, 2020

Bias mitigation - postprocessing

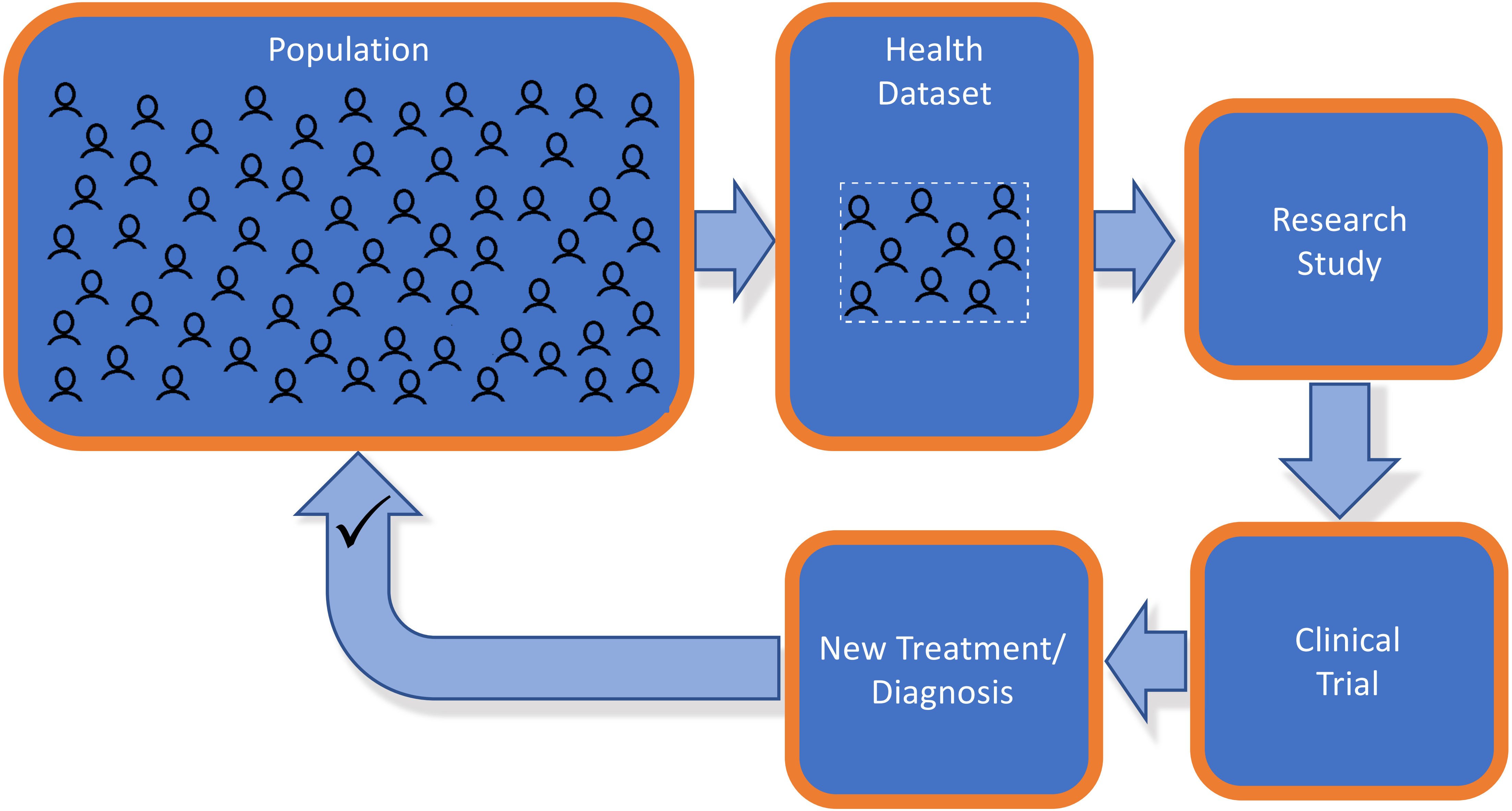
- *Adjust classification thresholds*:
  - For classification models, adjust thresholds to equalise FPR/TPR between protected groups (equalized odds) or just TPR (equal opportunity)



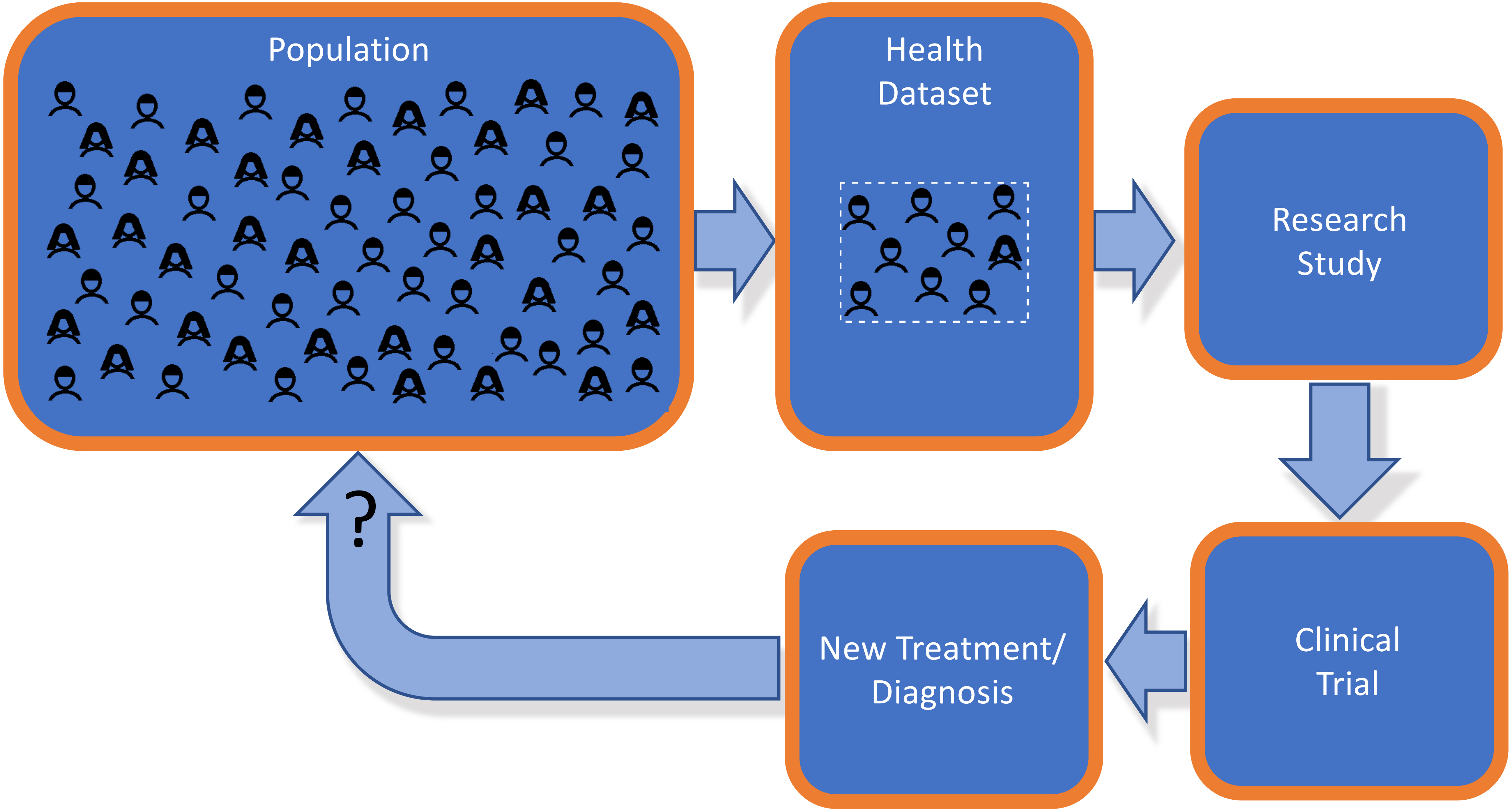
Hardt, et al, "Equality of Opportunity in Supervised Learning", Proceedings NeurIPS, 2016

Fair AI in medicine

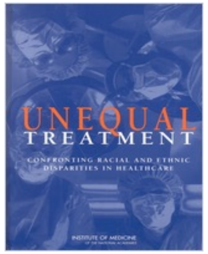
(Un)Fairness in modern medicine



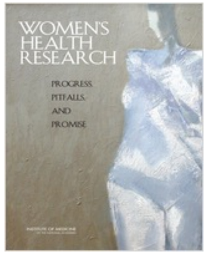
(Un)Fairness in modern medicine - sex



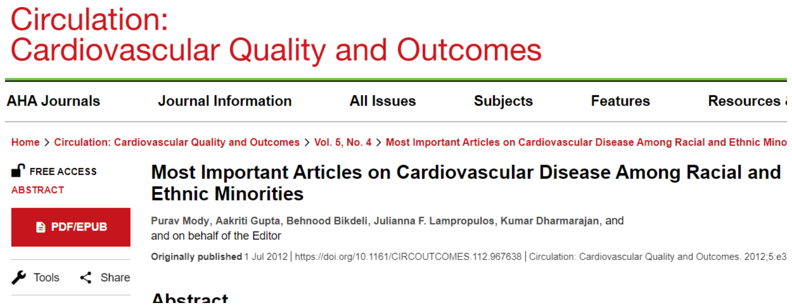
The bad news: Modern medicine is unfair



- "… race and ethnicity remain significant predictors of the quality of health care received …"



- "… medical research historically has neglected the health needs of women …"


Circulation: Cardiovascular Quality and Outcomes

AHA Journals | Journal Information | All Issues | Subjects | Features | Resources

Home > Circulation: Cardiovascular Quality and Outcomes > Vol. 5, No. 4 > Most Important Articles on Cardiovascular Disease Among Racial and Ethnic Mino

FREE ACCESS
ABSTRACT
PDF/EPUB
Tools
Share

Most Important Articles on Cardiovascular Disease Among Racial and Ethnic Minorities

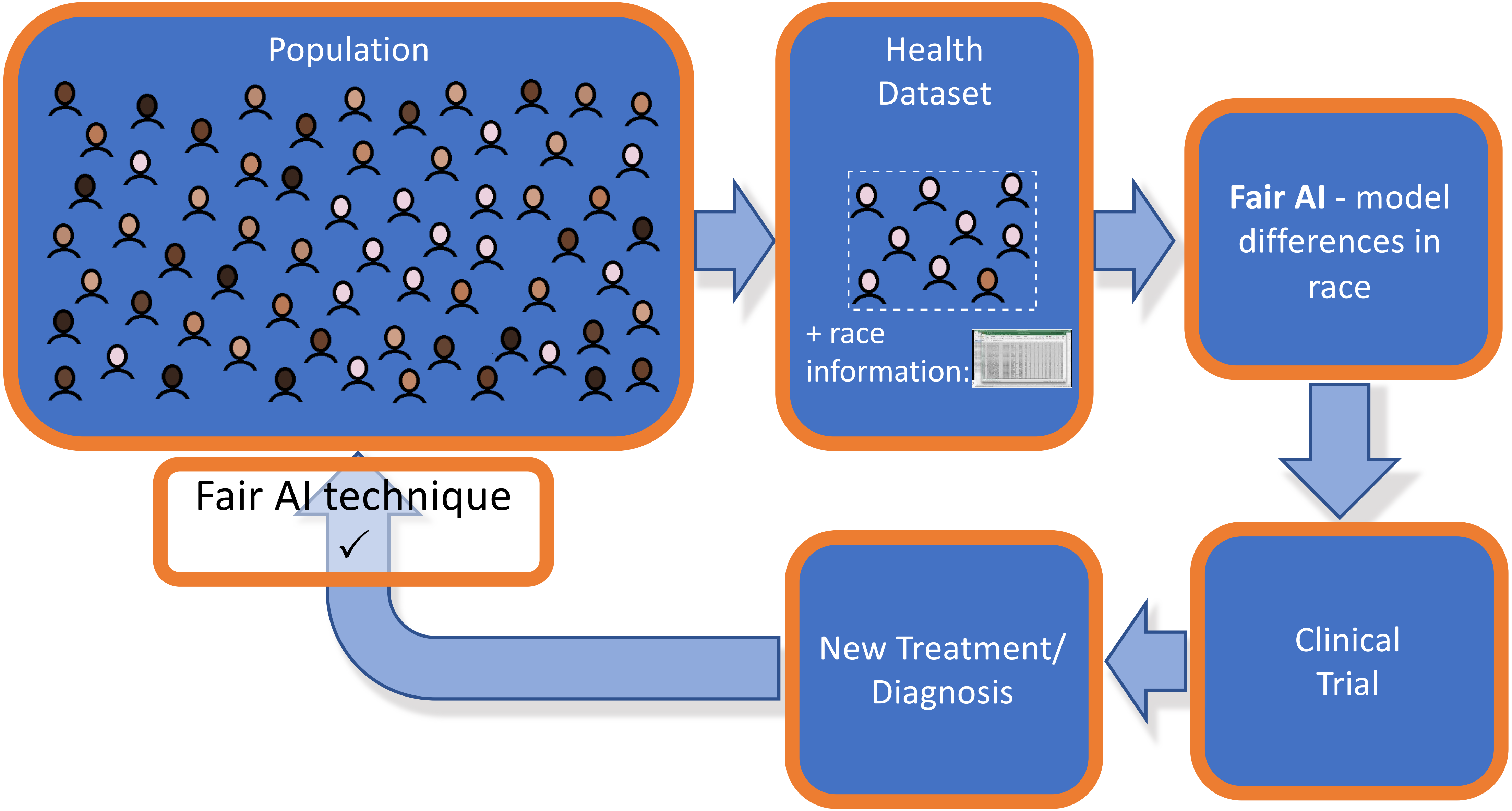
Purav Mody, Aakriti Gupta, Behnood Bikdeli, Julianna F. Lampropulos, Kumar Dharmarajan, and
and on behalf of the Editor

Originally published 1 Jul 2012 | https://doi.org/10.1161/CIRCOUTCOMES.112.967638 | Circulation: Cardiovascular Quality and Outcomes. 2012;5:e3

Abstract

- "Steady improvements have been made in the management of cardiovascular disease over the last decade. Yet, certain racial and ethnic minority groups have not experienced equivalent improvements in outcomes …"

The good news: Can AI be the solution?



A brief tour of fair AI in medicine

Literature on fair AI

	Bias assessment	Addressing bias
General machine learning/AI	100s or 1000s of papers	100s or 1000s of papers
Medical imaging	<50?	<10?

Bias assessment

Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis

Peter A. Noseworthy, MD[1,2], Zachi I. Attia, MSc[1], LaPrincess C. Brewer, MD, MPH[1], Sharonne N. Hayes, MD[1,3], Xiaoxi Yao, PhD[2,4], Suraj Kapa, MD[1], Paul A. Friedman, MD[1], Francisco Lopez-Jimenez, MD, MSc[1]

[1]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN

[2]Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN

[3]Office of Diversity and Inclusion, Mayo Clinic, Rochester, MN

[4]Division of Health Care Policy and Research, Department of Health Sciences Research, Mayo

"We recommend reporting of performance amongst diverse ethnic, racial, age and gender groups for all new AI tools to ensure responsible use of AI in medicine."

AI can identify race from medical images

2022:

Articles

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, Haoran Zhang

Summary

Background Previous studies in medical imaging have shown disparate abilities of artificial intelligence (AI) to detect a person's race, yet there is no known correlation for race on medical imaging that would be obvious to human experts when interpreting the images. We aimed to conduct a comprehensive evaluation of the ability of AI to recognise a patient's racial identity from medical images.

Methods Using private (Emory CXR, Emory Chest CT, Emory Cervical Spine, and Emory Mammogram) and public (MIMIC-CXR, CheXpert, National Lung Cancer Screening Trial, RSNA Pulmonary Embolism CT, and Digital Hand Atlas) datasets, we evaluated, first, performance quantification of deep learning models in detecting race from medical images, including the ability of these models to generalise to external environments and across multiple imaging modalities. Second, we assessed possible confounding of anatomic and phenotypic population features by assessing the ability of these hypothesised confounders to detect race in isolation using regression models, and by re-evaluating the deep learning models by testing them on datasets stratified by these hypothesised confounding variables. Last, by exploring the effect of image corruptions on model performance, we investigated the underlying mechanism by which AI models can recognise race.

Lancet Digit Health 2022; 4: e406–14

Published Online May 11, 2022 https://doi.org/10.1016/S2589-7500(22)00063-2

See Comment page e399

Department of Radiology (J W Gichoya MD, A R Bhimireddy MS, H Trivedi MD) and Department of Computer Science (Z Zaiman), Emory University, Atlanta, GA, USA; School of Computing, Informatics, and Decision Systems Engineering, Arizona State University,

Deep learning models can be trained to predict the race of the subject from multiple medical imaging datasets

Bias in skin lesion diagnosis

2020:

Risk of Training Diagnostic Algorithms on Data with Demographic Bias

Samaneh Abbasi-Sureshjani[1], Ralf Raumanns[1,2], Britt E. J. Michels[1], Gerard Schouten[1,2] and Veronika Cheplygina[1]

[1] Eindhoven University of Technology, The Netherlands
{s.abbasi, r.raumanns, v.cheplygina}@tue.nl
[2] Fontys University of Applied Science, The Netherlands
g.schouten@fontys.nl

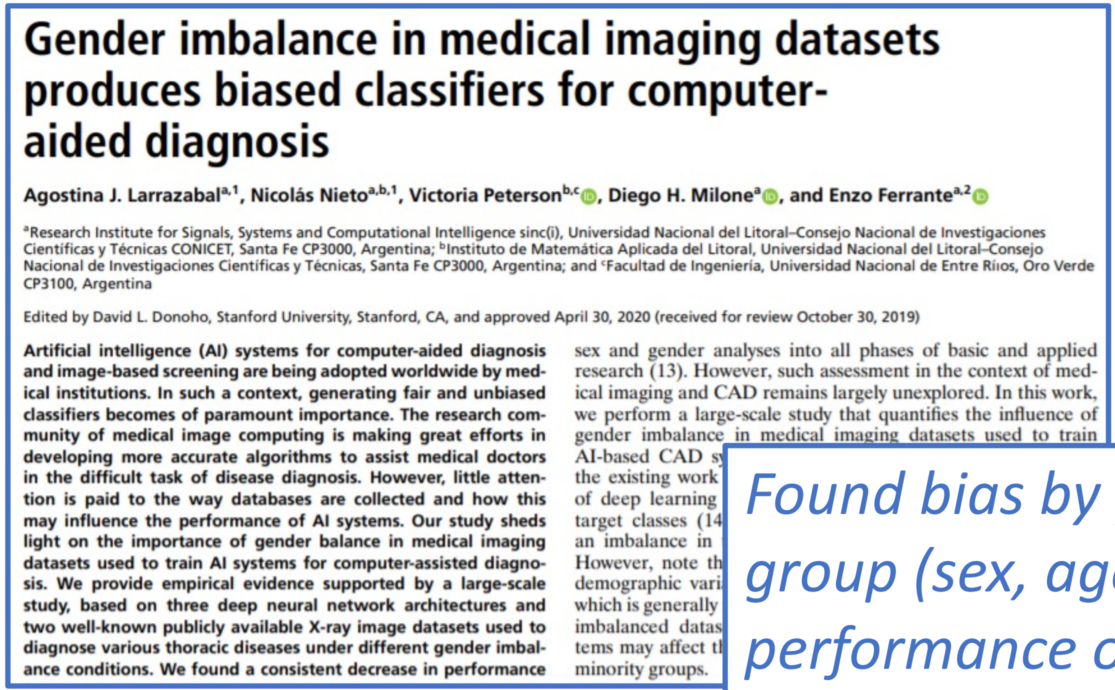
Abstract. One of the critical challenges in machine learning applications is to have fair predictions. There are numerous recent examples in various domains that convincingly show that algorithms trained with biased datasets can easily lead to erroneous or discriminatory conclusions. This is even more crucial in clinical applications where predictive algorithms are designed mainly based on a given set of medical images, and demographic variables such as age, sex and race are not taken into account. In this work, we conduct a survey of the MICCAI 2018 proceedings to investigate the common practice in medical image analysis applications. Surprisingly, we found that papers focusing on diagnosis rarely describe the demographics of the datasets used, and the diag-

Highlighted lack of reporting of demographic bias in the literature.

Found bias by protected group (sex, age) in CNN-based skin lesion image classifier, demonstrated simple bias mitigation scheme.

Bias in chest X-ray based AI

2020:



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

Agostina J. Larrazabal[a,1], Nicolás Nieto[a,b,1], Victoria Peterson[b,c], Diego H. Milone[a], and Enzo Ferrante[a,2]

[a]Research Institute for Signals, Systems and Computational Intelligence sinc(i), Universidad Nacional del Litoral–Consejo Nacional de Investigaciones Científicas y Técnicas CONICET, Santa Fe CP3000, Argentina; [b]Instituto de Matemática Aplicada del Litoral, Universidad Nacional del Litoral–Consejo Nacional de Investigaciones Científicas y Técnicas, Santa Fe CP3000, Argentina; and [c]Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Oro Verde CP3100, Argentina

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved April 30, 2020 (received for review October 30, 2019)

Artificial intelligence (AI) systems for computer-aided diagnosis and image-based screening are being adopted worldwide by medical institutions. In such a context, generating fair and unbiased classifiers becomes of paramount importance. The research community of medical image computing is making great efforts in developing more accurate algorithms to assist medical doctors in the difficult task of disease diagnosis. However, little attention is paid to the way databases are collected and how this may influence the performance of AI systems. Our study sheds light on the importance of gender balance in medical imaging datasets used to train AI systems for computer-assisted diagnosis. We provide empirical evidence supported by a large-scale study, based on three deep neural network architectures and two well-known publicly available X-ray image datasets used to diagnose various thoracic diseases under different gender imbalance conditions. We found a consistent decrease in performance

sex and gender analyses into all phases of basic and applied research (13). However, such assessment in the context of medical imaging and CAD remains largely unexplored. In this work, we perform a large-scale study that quantifies the influence of gender imbalance in medical imaging datasets used to train AI-based CAD sy the existing work of deep learning target classes (14 an imbalance in However, note th demographic vari which is generally imbalanced datas tems may affect th minority groups.

2021:

ARTICLES

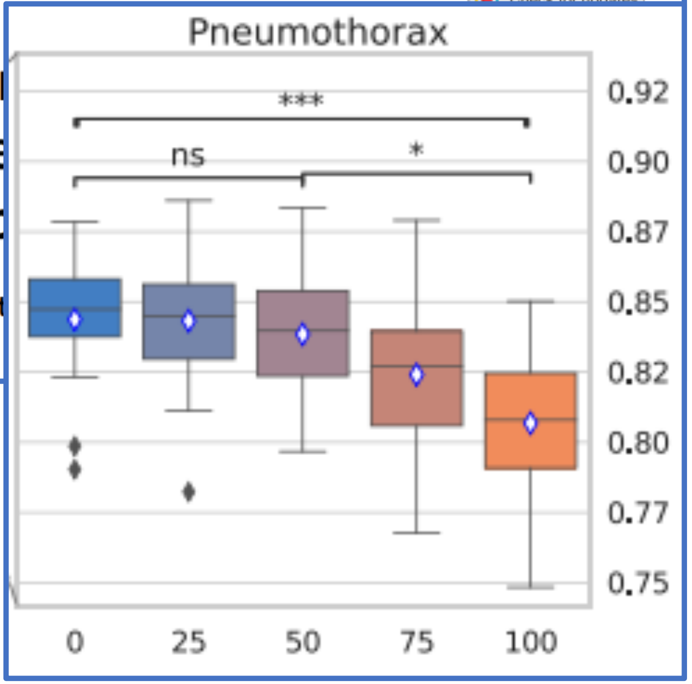
https://doi.org/10.1038/s41591-021-01595-0

nature medicine

OPEN

Underdiagnosis bias of a algorithms applied to che under-served patient pop

Haoran Zhang[3], Mat



Found bias by protected group (sex, age, race etc.) in performance of CNN-based chest-X-ray classifiers trained on public datasets.

Bias in brain imaging

> J Med Imaging (Bellingham). 2022 Nov;9(6):061102. doi: 10.1117/1.JMI.9.6.061102. Epub 2022 Aug 26.

Fairness-related performance and explainability effects in deep learning models for brain image analysis

Emma A M Stanley [1 2 3], Matthias Wilms [2 3 4], Pauline Mouches [1 2 3], Nils D Forkert [1 2 3 4]

Affiliations + expand

PMID: 36046104 PMCID: PMC9412191 (available on 2023-08-26)

Feature robustness and sex differences in medical imaging: a case study in MRI-based Alzheimer's disease detection

Eike Petersen[1 [0000-0003-0097-3868]], Aasa Feragen[1 [0000-0002-9945-981X]], Maria Luise da Costa Zemsch[1], Anders Henriksen[1], Oskar Eiler Wiese Christensen[1], and Melanie Ganz[2,3 [0000-0002-9120-8098]] for the Alzheimers Disease Neuroimaging Initiative*

[1] Technical University of Denmark, DTU Compute, Kgs. Lyngby, Denmark
Copenhagen, Department for Computer Science, Copenhagen, Denmark
ospitalet, Neurobiology Research Unit, Copenhagen, Denmark
ewipe@dtu.dk, afhar@dtu.dk, melanie.ganz@nru.dk

There may be bias in brain imaging tasks, but likely dependent on task …

Disproportionate Subgroup Impacts and C Challenges of Fairness in Artificial Intelligence for Medical Image Analysis

Emma A.M. Stanley[1,2,3 [0000-0002-7802-6820]], Matthias Wilms[2,3,4], and Nils D. Forkert[1,2,3,4 [0000-0003-2556-3224]]

[1] Department of Biomedical Engineering, University of Calgary, Canada
emma.stanley@ucalgary.ca
[2] Department of Radiology, University of Calgary, Canada
[3] Hotchkiss Brain Institute, University of Calgary, Canada
[4] Alberta Children's Hospital Research Institute, University of Calgary, Canada

OPEN ACCESS

Check for updates

Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies

Rongguang Wang[a,b], Pratik Chaudhari[a,c,1,2], and Christos Davatzikos[a,b,d,1,2,3]

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received July 18, 2022; accepted December 21, 2022

Despite the great promise that machine learning has offered in many fields of medicine, it has also raised concerns about potential biases and poor generalization across genders, age distributions, races and ethnicities, hospitals, and data acquisition equipment and protocols. In the current study, and in the context of three brain diseases, we provide evidence which suggests that when properly trained, machine learning models can generalize well across diverse conditions and do not necessarily suffer from bias. Specifically, by using multistudy magnetic resonance imaging consortia for diagnosing Alzheimer's disease, schizophrenia, and autism spectrum disorder, we find that well-trained models have a high area-under-the-curve (AUC) on subjects across different

Bias in image segmentation

Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation

Esther Puyol-Antón[1], Bram Ruijsink[1,2], Stefan K. Piechnik[7], Stefan Neubauer[7], Steffen E. Petersen[3,4,5,6], Reza Razavi[1,2], and Andrew P. King[1]

[1] School of Biomedical Engineering & Imaging Sciences, King's College London, UK
[2] Guy's and St Thomas' Hospital, London, UK.
[3] Willi
Queen
[4] Barts
[7] Divisio

A systematic study of race and sex bias in CNN-based cardiac MR segmentation

Tiarna Lee[1], Esther Puyol-Antón[1], Br
Andrew P.

[1] School of Biomedical Engineering & Imag
[2] Guy's and St Thomas'
[3] Department of Informatics,

A Study of Demographic Bias in CNN-based Brain MR Segmentation

Stefanos Ioannou[1], Hana Chockler[3,1], Alexander Hammers[2], and Andrew P.
nitiative*
n, U.K.
iences,

Fairness in Cardiac Magnetic Resonance Imaging: Assessing Se
and Racial Bias in Deep Learning-Based Segmentation

Esther Puyol-Antón[1], Bram Ruijsink[1,2,3], Jorge Mariscal Harana[1], Stefan K. Piechn
Stefan Neubauer[4], Steffen E. Petersen[5,6,7,8], Reza Razavi[1,2], Phil Chowienczyk[1,9] and
Andrew P. King[1]*

*[1] School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom, [2] De
of Adult and Paediatric Cardiology, Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom, [3] Divi
of Heart and Lungs, Department of Cardiology, University Medical Centre Utrecht, Utrecht, Netherlands, [4] Division
of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom, [5] Na
Institute for Health Research (NIHR) Barts Biomedical Research Centre, William Harvey Research Institute, Queen
University London, London, United Kingdom, [6] Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS
London, United Kingdom, [7] Health Data Research UK, London, United Kingdom, [8] Alan Turing Institute, London,
United Kingdom, [9] British Heart Foundation Centre, King's College London, London, United Kingdom*

ELSEVIER
Contents lists available at ScienceDirect
Computer Methods and Programs in Biomedicine
journal homepage: www.elsevier.com/locate/cmpb

Understanding skin color bias in deep learning-based skin lesion segmentation

Marin Benčević[a,b,*], Marija Habijan[a], Irena Galić[a], Danilo Babin[c], Aleksandra Pižurica[b]

[a] J. J. Strossmayer University of Osijek, Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Kneza Trpimira 2B, Osijek, 31000, Croatia
[b] Ghent University, Department of Telecommunications and Information Processing, TELIN-GAIM, St-Pietersnieuwstraat 41, Ghent, 9000, Belgium
[c] Ghent University, Department of Telecommunications and Information Processing, imec-TELIN-IPI, St-Pietersnieuwstraat 41, Ghent, 9000, Belgium

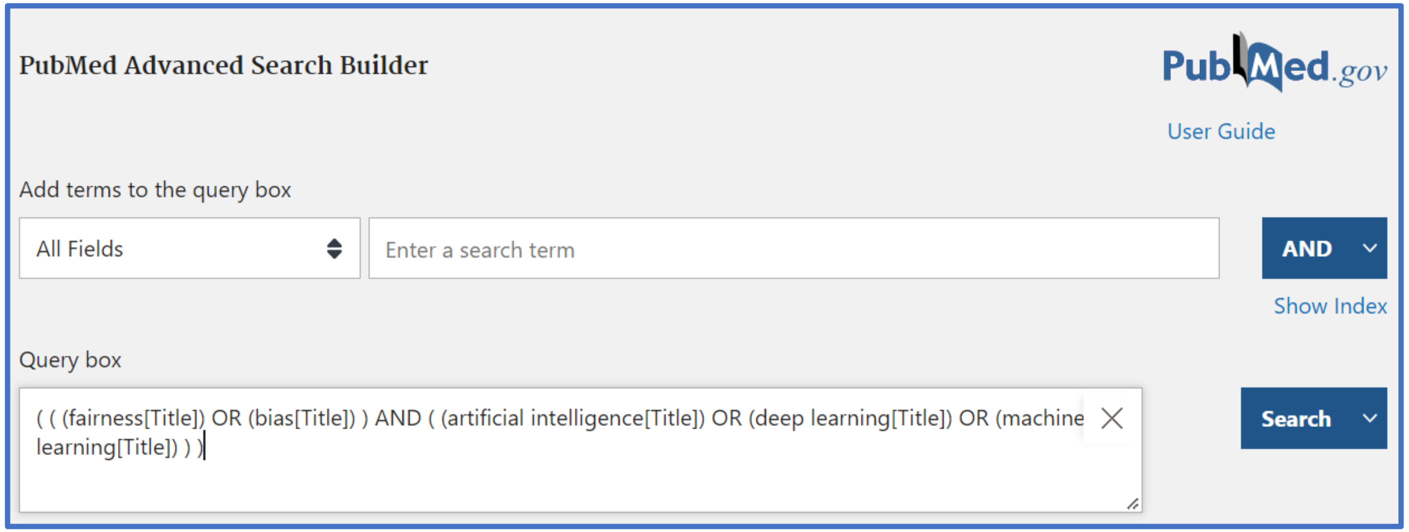
A R T I C L E I N F O

Keywords:
AI fairness
Dermatological image analysis
Deep neural networks
Skin lesion segmentation

A B S T R A C T

Background: The field of dermatological image analysis using deep neural networks includes the semantic segmentation of skin lesions, pivotal for lesion analysis, pathology inference, and diagnoses. While biases in neural network-based dermatoscopic image classification against darker skin tones due to dataset imbalance and contrast disparities are acknowledged, a comprehensive exploration of skin color bias in lesion segmentation models is lacking. It is imperative to address and understand the biases in these models.

Fairness of AI in Medical Imaging – a growing field!



What is (and isn't) Fair AI?

Fairness and domain shift

- *Domain shift* – a change in the distribution between a source dataset and a target dataset
- In AI, often source=*training* and target=*test* …

Examples of domain shift

Computer vision

- E.g.

Office-Home dataset

PACS dataset

Medical imaging

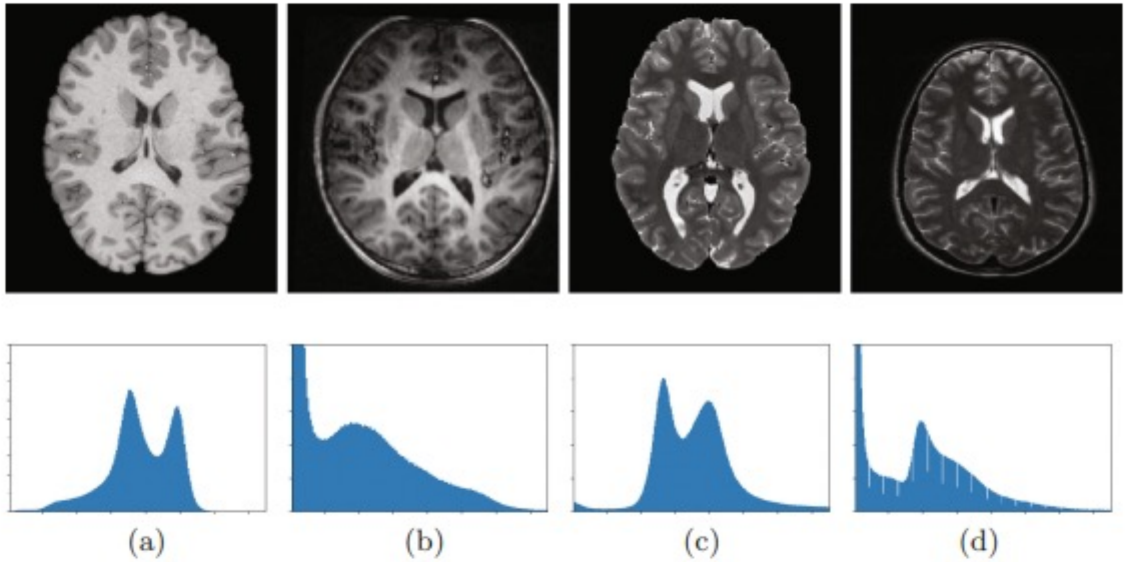


Fig. 1. Image slices (top) and corresponding histograms (bottom) of normalized T1w (a,b) and T2w (c,d) MRIs from different scanners. Despite high-level information similarity, there exists considerable intensity and contrast differences, which segmentation algorithms are often sensitive to.

https://paperswithcode.com/dataset/office-home

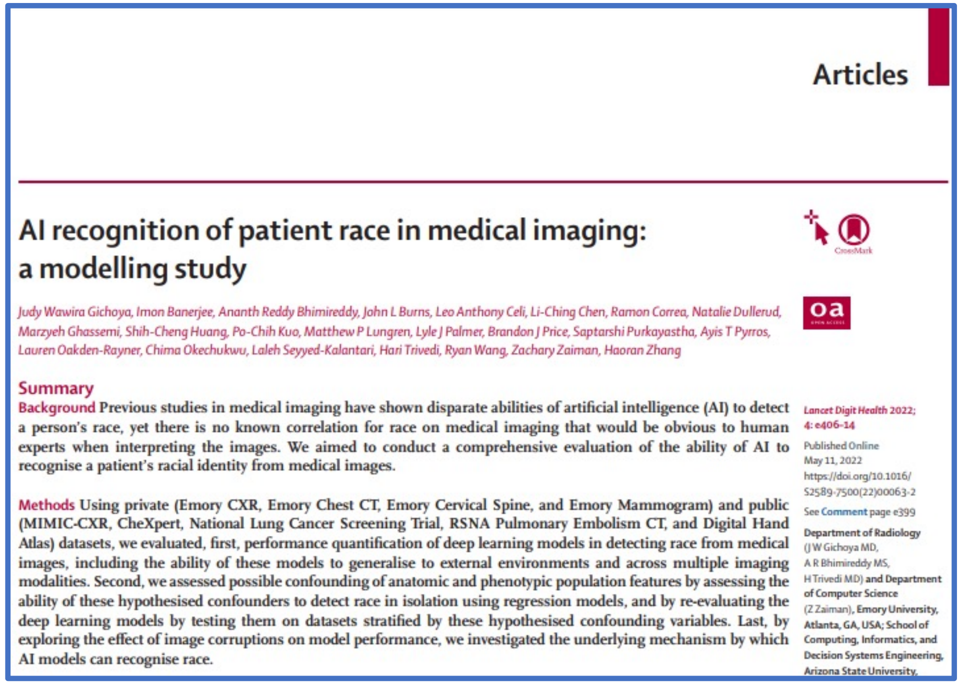
https://paperswithcode.com/dataset/pacs

N. Karani, et al. (2018) "A Lifelong Learning Approach to Brain MR Segmentation Across Scanners and Protocols", Proceedings MICCAI.

So is fair AI just demographic domain shift?

- A number of works have found similar domain shifts between different *protected groups*, e.g.

Race



Articles

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, Haoran Zhang

Summary

Background Previous studies in medical imaging have shown disparate abilities of artificial intelligence (AI) to detect a person's race, yet there is no known correlation for race on medical imaging that would be obvious to human experts when interpreting the images. We aimed to conduct a comprehensive evaluation of the ability of AI to recognise a patient's racial identity from medical images.

Methods Using private (Emory CXR, Emory Chest CT, Emory Cervical Spine, and Emory Mammogram) and public (MIMIC-CXR, CheXpert, National Lung Cancer Screening Trial, RSNA Pulmonary Embolism CT, and Digital Hand Atlas) datasets, we evaluated, first, performance quantification of deep learning models in detecting race from medical images, including the ability of these models to generalise to external environments and across multiple imaging modalities. Second, we assessed possible confounding of anatomic and phenotypic population features by assessing the ability of these hypothesised confounders to detect race in isolation using regression models, and by re-evaluating the deep learning models by testing them on datasets stratified by these hypothesised confounding variables. Last, by exploring the effect of image corruptions on model performance, we investigated the underlying mechanism by which AI models can recognise race.

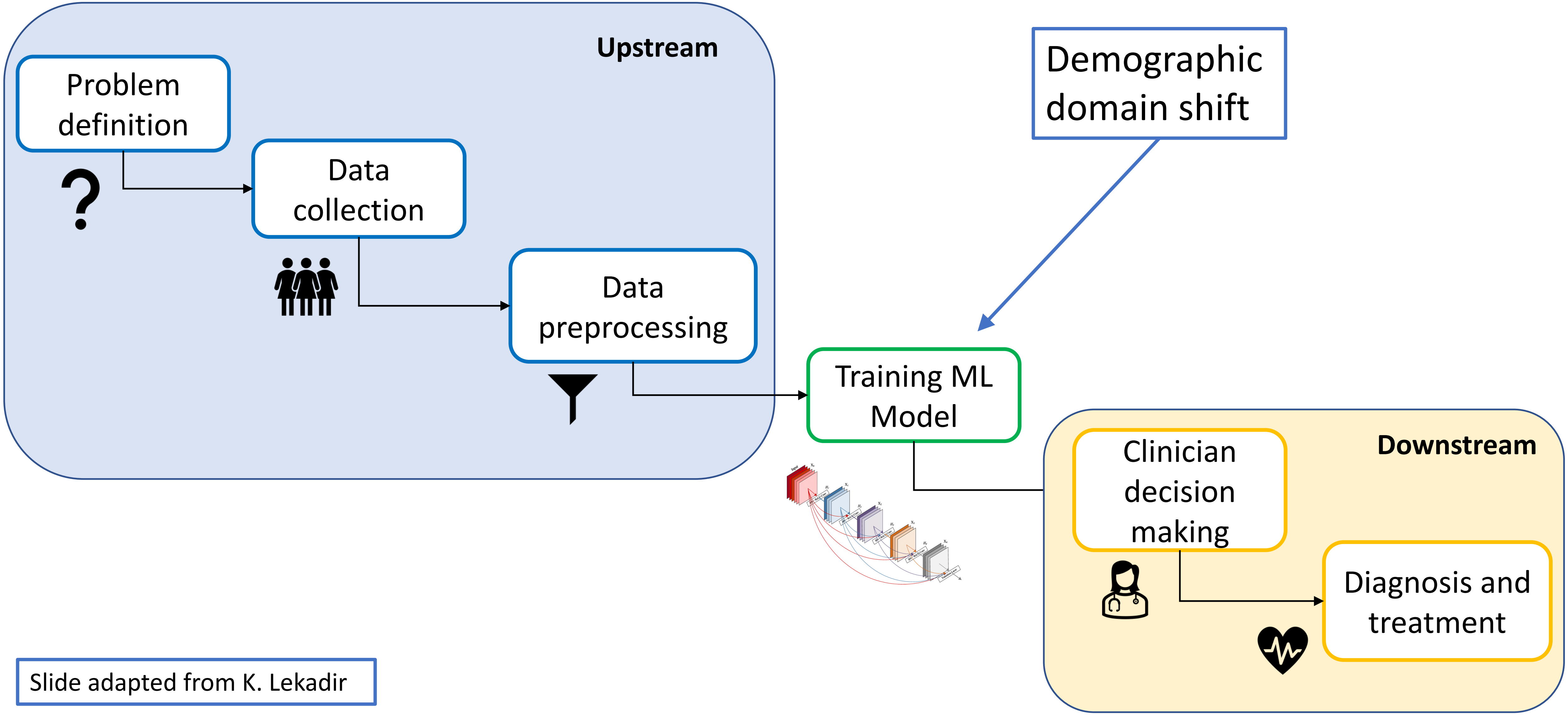
Lancet Digit Health 2022; 4: e406–14

Published Online May 11, 2022 https://doi.org/10.1016/S2589-7500(22)00063-2

See Comment page e399

Department of Radiology (J W Gichoya MD, A R Bhimireddy MS, H Trivedi MD) and Department of Computer Science (Z Zaiman), Emory University, Atlanta, GA, USA; School of Computing, Informatics, and Decision Systems Engineering, Arizona State University,

Well yes, but ...



Slide adapted from K. Lekadir

Fair AI in computer vision and medical imaging

Computer vision

E.g. *facial recognition AI*:
- Nothing known about subject before inference
- Seems uncontroversial that we should minimise differences in recognition accuracy
- *Bias mitigation*

Medical imaging

E.g. *AI-assisted diagnosis*:
- We know the patient's sex, race & clinical history
- Mitigating bias might adversely affect some protected groups
- *Do no harm …*

Non-maleficence

- The ethical principle of doing no harm, expressed in the ancient medical maxim *primum non nocere* (first do no harm)
- In demographic domain shift, there are important *ethical considerations* when assessing/mitigating bias, e.g.

Cardiac MR segmentation results: (Dice coefficients, high=good)

But when we try to reduce the bias, performance for White subjects gets worse!

Race bias in performance!

-> **Performance/fairness trade-off** …

Approach	Segmentation							Fairness	
	White	Mixed	Asian	Black	Chinese	Others	Avg	SD	SER
1. Fairness through unawareness	93.51	84.52	88.90	85.88	87.63	85.66	87.68	**3.25**	**2.38**
2. Stratified batch sampling	90.88	93.84	93.65	93.07	94.35	93.50	93.22	**1.22**	**1.62**
3. Fair meta-learning for segmentation	92.75	88.03	90.64	89.60	88.18	88.27	89.58	**1.86**	**1.81**
4. Protected group models	91.03	93.17	93.34	92.15	93.04	93.08	92.64	**0.89**	**1.35**

www.oxfordreference.com

E. Puyol-Anton, et al. (2022) "Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation", Front Cardiovasc Med.

What does fairness mean in a medical context?

2021 paper:

Assessing Bias in Medical AI

Melanie Ganz [1 2] Sune H. Holm [3] Aasa Feragen [4 2]

Abstract

Machine learning and artificial intelligence are increasingly deployed in critical societal functions such as finance, media and healthcare. Along with their deployment come increasing reports

bias and promoting fairness in medical AI.

2. Case discussions

In the two cases discussed below, bias has different sources,

2022 paper:

Comment

https://doi.org/10.1038/s41467-022-32186-3

Addressing fairness in artificial intelligence for medical imaging

María Agustina Ricci Lara, Rodrigo Echeveste and Enzo Ferrante

Check for updates

A plethora of work has shown that AI systems can systematically and unfairly be biased against certain populations in multiple scenarios. The field of medical imaging, where AI systems are beginning to be increasingly adopted, is no exception. Here we discuss the meaning of fair-

broad field of healthcare have recently been surveyed and discussed[7], in this comment we focus on the sub-field of medical imaging. Indeed, when it comes to biases in ML systems that can benefit certain subpopulations in detriment of others, the field of medical imaging is not the exception[8,9]. In what follows we will comment on recent work in the field and highlight valuable unexplored areas of research, discussing potential challenges and available strategies.

2022 blog:

What do we want from fair AI in medical imaging?

January 11, 2022 / 1 Comment / Andrew King

www.kclmmag.org/blog/

2022 RadioNuclear podcast:

Summary & resources

Summary

- *Fair AI* is the analysis of *biased* or *toxic* behaviour in AI models, and the development of techniques to *address* this behaviour

- *Bias* should be *addressed* in *all parts of the healthcare AI pipeline*, including defining the problem, identifying and collecting the data etc.

- Lots of areas for future work in Fair AI for healthcare!

Conclusions …

- When implementing AI for medical imaging:
  - *Important to consider “if” and “how” rather than just “can”*
  - *Ensure as far as possible that benefits of AI are shared equitably*
  - *Be open and transparent about limits of performance*

Resources/toolkits for fair AI

- AI Fairness 360 - https://aif360.mybluemix.net/
  - Created by IBM but open-source
- Fairlearn - https://fairlearn.org/
  - Open source, originally Microsoft, self-governing from 2021 - https://fairlearn.org/v0.8/about/index.html
- Aequitas - http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/
  - Open source bias audit toolkit for machine learning developers
- MEDFAIR - https://github.com/ys-zong/MEDFAIR
  - Fairness benchmarking suite for medical imaging

Fairness initiatives

- FUTURE-AI - https://future-ai.eu/
  - Establish guidelines for AI in healthcare, including fairness as key principle
- STANDING TOGETHER - https://www.datadiversity.org/
  - Promotes the formation of inclusive, diverse and transparent medical datasets
- FAIMI - https://faimi-workshop.github.io/
  - Independent academic initiative aimed at exploring and promoting Fair AI in medical imaging

Research on fairness, equity, and accountability in the context of machine learning has extensively demonstrated ethical risks in the deployment of such systems, including for medical image analysis. In a series of interdisciplinary events, our aim is to advance the discourse around fairness issues in the medical image analysis community.

Next up: ISBI 2024 tutorial Fairness of AI in Medical Imaging (FAIMI)!

Organizers

Aasa Feragen, DTU Compute, Technical University of Denmark
Andrew King, King's College London
Ben Glocker, Imperial College London
Daniel Moyer, Vanderbilt University
Enzo Ferrante, CONICET, Universidad Nacional del Litoral
Eike Petersen, Fraunhofer Institute for Digital Medicine MEVIS, Germany
Esther Puyol-Antón, HeartFlow and King's College London
Melanie Ganz-Benjaminsen, University of Copenhagen & Neurobiology Research Unit, Rigshospitalet
Veronika Cheplygina, IT University Copenhagen

Two events:

MICCAI 2024 workshop
(in-person, Morocco)

+

Free online symposium
(in November 2024)