An Empirical Framework for Domain Generalization in Clinical Settings

Haoran Zhang[1,2], Natalie Dullerud[1,2], Laleh Seyyed-Kalantari[1,2], Quaid Morris[3], Shalmali Joshi[4], Marzyeh Ghassemi[1,2]

[1]University of Toronto [2]Vector Institute [3]MSKCC [4]Harvard University

Our Contributions

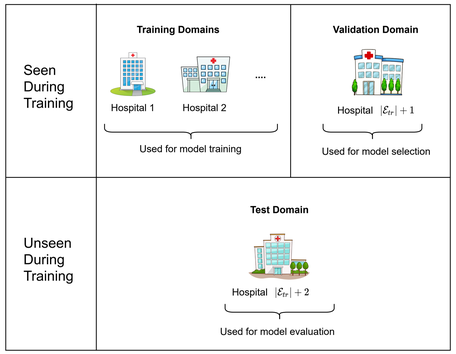
We propose an experimental framework for benchmarking domain generalization methods on clinical data.

- We show that state of the art domain generalizations do not perform significantly better than ERM on real-world clinical imaging data.
- We introduce a framework which generates plausible augmented versions of clinical datasets with domain shift.
- We find there exist limited cases of synthetic shift where domain generalization methods perform better than ERM – only when the spurious correlation is extreme.

Paper: https://arxiv.org/abs/2103.11163
Code: https://github.com/MLforHealth/ClinicalDG

What is Domain Generalization?

- A learning setup where you are given labelled data from multiple training environments.
- The goal is to learn a predictor that can perform well on all possible environments.



- No access to any data from the target environment during training.

Setup	Training inputs	Test inputs
Generative learning	U^1	$\emptyset$
Unsupervised learning	U^1	U^1
Supervised learning	L^1	U^1
Semi-supervised learning	L^1, U^1	U^1
Multitask learning	$L^1, \ldots, L^{d_{tr}}$	$U^1, \ldots, U^{d_{tr}}$
Continual (or lifelong) learning	$L^1, \ldots, L^\infty$	$U^1, \ldots, U^\infty$
Domain adaptation	$L^1, \ldots, L^{d_{tr}}, U^{d_{tr}+1}$	$U^{d_{tr}+1}$
Transfer learning	$U^1, \ldots, U^{d_{tr}}, L^{d_{tr}+1}$	$U^{d_{tr}+1}$
Domain generalization	$L^1, \ldots, L^{d_{tr}}$	$U^{d_{tr}+1}$

Reproduced from Gulrajani and Lopez-Paz, 2020

Prior Work

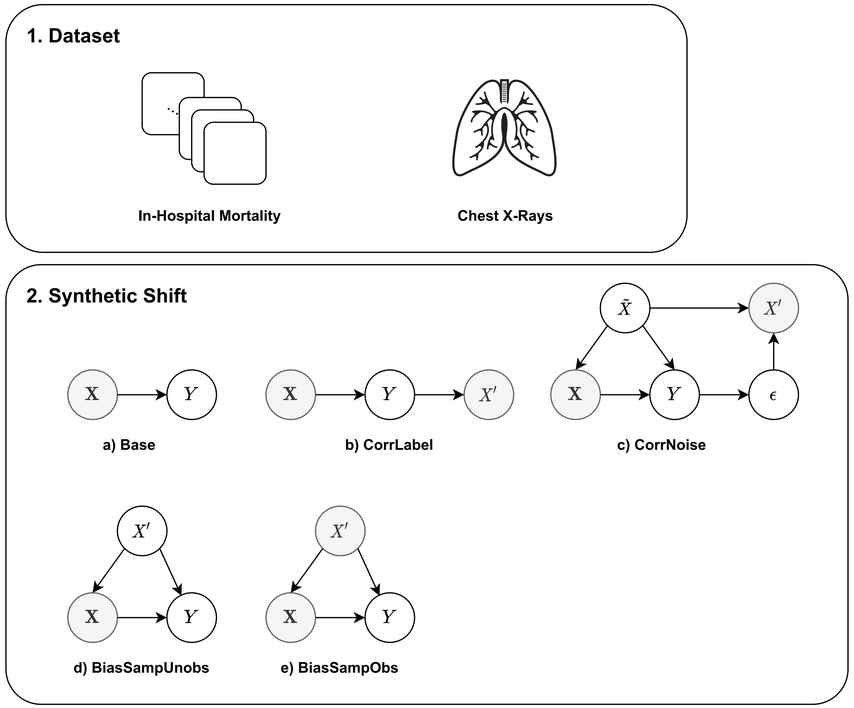
Domain Generalization Methods

- Empirical Risk Minimization (ERM): Train on pooled data from training environments
- Invariance learning: IRM (Arjovsky et al., 2019), etc
- Robust Optimization: VREx (Krueger et al., 2020), etc
- ... and many more.

Domain Generalization Benchmarks

- DomainBed (Gulrajani and Lopez-Paz, 2020) and WILDS (Koh et al., 2020)
- Domain generalization methods don't out-perform ERM on general imaging benchmarks!

Experimental Framework



Datasets

	In-Hospital Mortality (eICU)					Chest X-Rays (CXR)			
Environment	**Midwest**	**West**	**Northeast**	**Missing**	**South**	**MIMIC-CXR**	**CheXpert**	**Chest-Xray8**	**PadChest**
Assigned Split	Train	Train	Train	Validation	Test	Train	Train	Validation	Test
# Samples	10,985	4,527	2,495	1,846	10,827	249,995	191,229	112,120	99,934
% Positive	9.43%	14.42%	13.19%	12.68%	11.74%	7.37%	2.45%	1.28%	4.90%

Base Results

- AUROCs on test environment

Model Selection	Training Domains			Validation Domains		
Dataset	**eICU**	**CXR (multitask)**	**CXR (binary)**	**eICU**	**CXR (multitask)**	**CXR (binary)**
Oracle	0.856±0.017	0.882±0.007	0.810±0.036	0.856±0.017	0.882±0.007	0.810±0.036
ERM	0.867±0.002	**0.850±0.008**	**0.721±0.057**	0.868±0.003	0.840±0.014	**0.723±0.045**
GroupDRO	0.864±0.008	0.844±0.015	0.720±0.076	0.864±0.004	0.840±0.009	0.717±0.021
IRM	0.868±0.005	0.811±0.071	0.597±0.075	0.866±0.002	0.809±0.052	0.600±0.090
VREx	**0.869±0.004**	0.845±0.029	0.671±0.122	0.866±0.003	0.833±0.013	0.686±0.051
RVP	0.868±0.004	0.765±0.101	0.563±0.156	**0.869±0.003**	0.754±0.125	0.626±0.088
IGA	0.815±0.072	0.770±0.028	0.574±0.061	0.825±0.064	0.768±0.028	0.553±0.032
CORAL	0.866±0.007	0.845±0.010	0.709±0.103	0.867±0.004	**0.849±0.015**	0.689±0.095
MLDG	0.867±0.004	0.779±0.028	0.486±0.046	0.861±0.005	0.774±0.026	0.603±0.047

Biased Subsampling Shift

- Define desired data parameters $\mu^e_M = P(Y=1|G=M)$, $\mu^e_F = P(Y=1|G=F)$
- Subsample genders separately until desired label prevalences are reached – gender is now a confounder.

Dataset	Environment	μ_M	μ_F	% Male	% Female
eICU	Midwest	0.8	0.05	35.7%	64.3%
	West	0.7	0.1	57.6%	42.4%
	Northeast	0.6	0.15	51.2%	48.8%
	Missing	0.3	0.3	50.3%	49.7%
	South	0.1	0.5	82.8%	17.2%
CXR	MIMIC-CXR	0.2	0.02	30.2%	69.8%
	CheXpert	0.1	0.03	28.6%	71.4%
	ChestX-ray8	0.07	0.04	30.8%	69.2%
	PadChest	0.05	0.05	54.6%	45.4%

Biased Subsampling Shift Results

- In-Hospital Mortality - AUROCs on test environment

Model Selection	Training Domains		Validation Domain	
Observed	**No**	**Yes**	**No**	**Yes**
Oracle	0.883±0.015	0.891±0.009	0.883±0.015	0.891±0.009
ERM	0.759±0.011	0.645±0.018	0.791±0.016	**0.721±0.017**
GroupDRO	**0.767±0.015**	0.638±0.019	**0.793±0.008**	0.712±0.016
IRM	**0.767±0.015**	0.639±0.023	0.790±0.012	0.709±0.041
VREx	0.760±0.009	0.622±0.065	0.786±0.016	0.699±0.023
RVP	0.761±0.021	0.649±0.025	0.783±0.017	0.695±0.017
IGA	0.749±0.041	0.554±0.090	0.737±0.054	0.569±0.091
CORAL	0.762±0.009	**0.671±0.012**	0.776±0.026	0.710±0.015
MLDG	0.759±0.023	0.643±0.023	0.789±0.010	0.686±0.037

- Chest X-Rays - AUROCs on test environment

Model Selection	Training Domains		Validation Domain	
Observed	**No**	**Yes**	**No**	**Yes**
Oracle	0.820±0.029	0.812±0.019	0.820±0.029	0.812±0.019
ERM	**0.640±0.028**	0.618±0.088	**0.669±0.048**	**0.662±0.043**
GroupDRO	0.628±0.035	0.590±0.077	0.664±0.041	0.617±0.041
IRM	0.575±0.067	0.506±0.063	0.608±0.048	0.608±0.070
VREx	0.614±0.075	0.595±0.052	0.629±0.047	0.644±0.066
RVP	0.557±0.060	0.544±0.072	0.575±0.116	0.617±0.068
IGA	0.579±0.040	0.542±0.061	0.611±0.039	0.635±0.037
CORAL	0.623±0.045	**0.651±0.043**	0.649±0.023	0.652±0.018
MLDG	0.513±0.030	0.563±0.015	0.609±0.036	0.604±0.026

References

- Arjovsky, Martin, et al. "Invariant risk minimization." (2019).
- Gulrajani, Ishaan, and David Lopez-Paz. "In search of lost domain generalization" (2020).
- Koh, Pang Wei, et al. "WILDS: A Benchmark of in-the-Wild Distribution Shifts." (2020).
- Krueger, David, et al. "Out-of-Distribution Generalization via Risk Extrapolation (REx)" (2020).