Population-aware Hierarchical Bayesian Domain Adaptation

Vishwali Mhasawade[1], Nabeel Abdur Rehman[1], Rumi Chunara[1,2]

[1]Computer Science & Engineering, Tandon School of Engineering [2]Biostatistics, College of Global Public Health

New York University

ABSTRACT

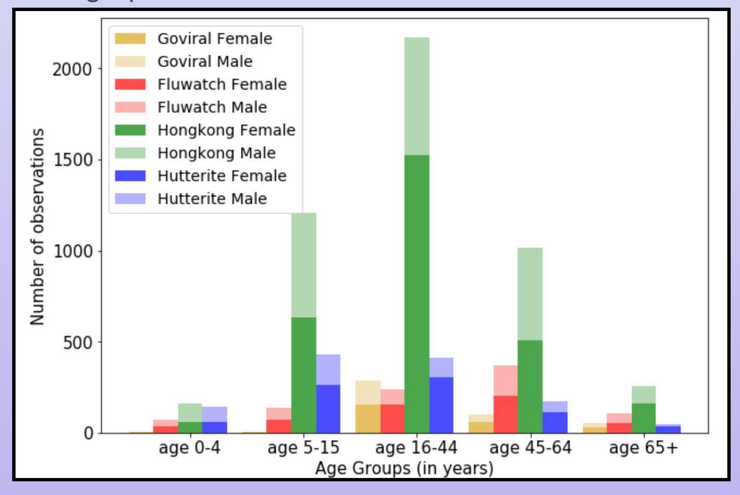
Population attributes are essential in health for understanding who the data represents and precision medicine efforts [1]. Even within disease infection labels, patients can exhibit significant variability; “fever" may mean something different when reported in a doctor's office versus from an online app, precluding directly learning across different data sets for the same prediction task. This problem falls into the domain adaptation paradigm [2]. However, research in this area has to-date not considered who generates the data; symptoms reported by a woman versus a man, for example, could also have different implications. We propose a novel population-aware domain adaptation approach by formulating the domain adaptation task as a multi-source hierarchical Bayesian framework. The model improves prediction in the case of largely unlabeled target data by harnessing both domain and population invariant information.

DATA

TABLE 1: Summary of our datasets (real-world, from two different domains: Citizen science and Health-worker facilitated, source references in paper)

Study	Location	Observations (positive)	Collection Type
Goviral	New York City	520 (291)	Citizen Science
Fluwatch	United Kingdom	915 (567)	Citizen Science
Hongkong	Hong Kong	4954 (1471)	Health worker Facilitated
Hutterite	Alberta, Canada	1281 (787)	Health worker Facilitated

FIGURE 1: Demographic distribution across datasets



METHOD

1. Learn all the parameters jointly from the source and partially labelled target data by a Hierarchical Bayesian approach (objective below, θ_j^d denotes the parameter for symptom j in domain $\mathcal{D}$, θ_k^d, the symptom k in a particular dataset d, f_j is a statistical measure of symptom j in the dataset, β is the weight (influence balance between node and parent), λ is a regularizing parameter, Div($\theta^l, \theta^{par(c)}$) is a divergence over child and parent parameters [3].
2. Learn weights of each level of the hierarchy for the target dataset (Logistic Regression).

$$F_{objective} = -\sum_{d\,\epsilon\mathcal{D}}[\sum_{j}(f_j + \lambda) + \theta_j^d - \log\sum_{k}\exp(\theta_k^d)] + \beta \sum_{l\,\epsilon\,Nodes} Div(\theta^l, \theta^{par(c)})$$

FIGURE 2: Population-aware Hierarchical model - θ: parameters at different nodes, $\mathcal{D}$: different domains, α: the priors. (A) Root level that represents invariant information across all data, (B) population parameters and information invariant to population-attributes (here, age and gender), (C) dataset and domain-specific parameters and information (here, for citizen science (CS) and health-worker facilitated (HW) domain datasets).

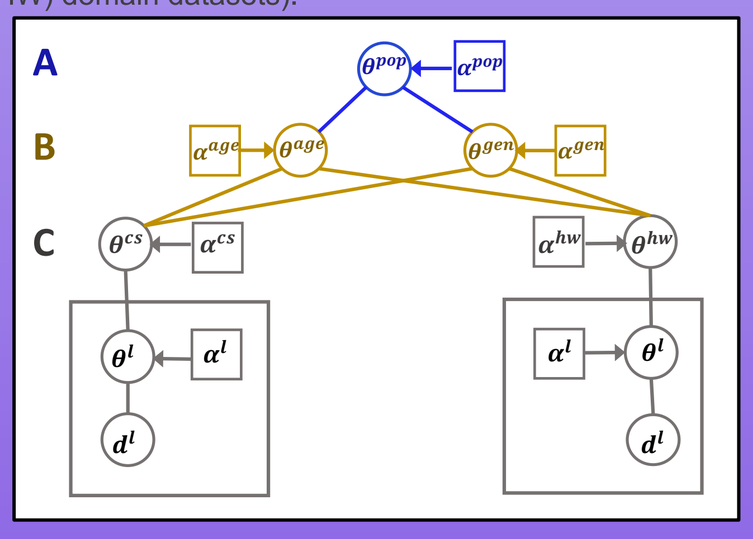
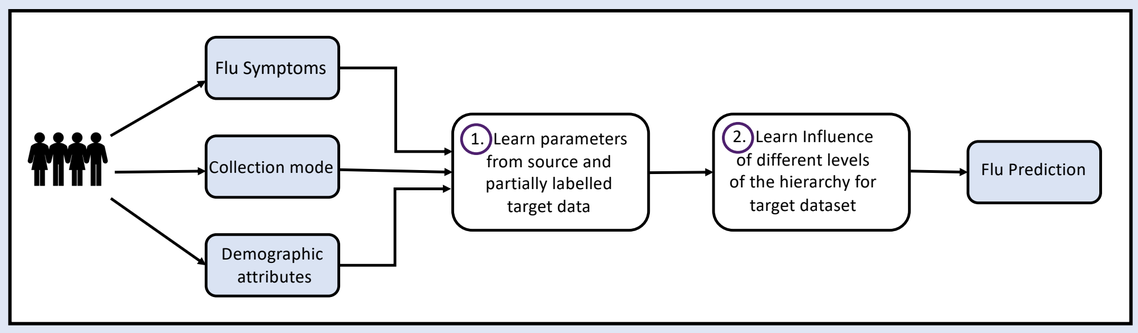


FIGURE 3: Overall approach.



RESULTS

We evaluate compared to baselines that assess incorporation of population attributes as well as a hierarchical approach.

TABLE 2: AUC results for flu prediction task (with 20% labelled data from the target dataset).

Method \ Dataset	Goviral	Fluwatch	Hongkong	Hutterite
Target Only	0.652	0.590	0.890	0.749
Logistic Regression	0.681	0.461	0.882	0.748
FEDA (Only symptoms) [4]	0.675	0.521	0.900	0.726
FEDA+p(With demographics)	0.693	0.612	0.914	0.824
Hierarchical (Only symptoms)	0.685	0.486	0.889	0.719
Hierarchical+p (With demographics)	0.710	0.627	0.918	0.827

Different levels in the hierarchy have different information control for each dataset.

- In the Goviral dataset, for example, gender parameters are close to each other since the proportion of positive flu males and females is close to 1.
- Overall, the top-performing models utilize demographic attributes → population attributes contain invariant information that can be harnessed when labelling data infection data.

TABLE 3: AUC scores across increasing proportions of training data (best two models)

Proportion	10%		15%		20%		25%	
Dataset	FEDA+p	Hier+p	FEDA+p	Hier+p	FEDA+p	Hier+p	FEDA+p	Hier+p
Goviral	0.670	0.671	0.634	0.664	0.693	0.71	0.664	0.690
Fluwatch	0.724	0.576	0.718	0.699	0.613	0.627	0.757	0.710
Honkong	0.896	0.911	0.971	0.984	0.914	0.918	0.969	0.940
Hutterite	0.742	0.785	0.873	0.880	0.824	0.827	0.879	0.800

CONCLUSION

The Population-aware Hierarchical Bayesian Domain Adaptation method outperforms existing domain adaptation methods for predicting influenza infection in situations of low-amounts of labelled data (which can happen commonly as laboratory tests are expensive). The method incorporates learning from multiple sources and population-invariant information. Given these findings, we are interested in developing a generalizable framework for understanding how domain and population distribution differences affect results, and how this can be used in any situation with non-representative samples.

REFERENCES

[1] Ray, Bisakha, and Rumi Chunara. "Predicting Acute Respiratory Infections from Participatory Data." *Online journal of public health informatics* 9, no. 1 (2017).

[2] Rehman, Nabeel Abdur, Maxwell Matthaios Aliapoulios, Disha Umarwani, and Rumi Chunara. "Domain Adaptation for Infection Prediction from Symptoms Based on Data from Different Study Designs and Contexts." *arXiv preprint arXiv:1806.08835* (2018).

[3] Elidan, Gal, Ben Packer, Geremy Heitz, and Daphne Koller. "Convex point estimation using undirected bayesian transfer hierarchies." *arXiv preprint arXiv:1206.3252* (2012).

[4] Daumé III, Hal. "Frustratingly easy domain adaptation." *arXiv preprint arXiv:0907.1815* (2009).

ACKNOWLEDGEMENTS:

This project was supported in part by grants from the National Science Foundation (1643576, 1551036).