Population-aware Hierarchical Bayesian Domain Adaptation via Multi-component Invariant Learning

Vishwali Mhasawade
vishwalim@nyu.edu
New York University

Nabeel Abdur Rehman
nabeel@nyu.edu
New York University

Rumi Chunara
rumi.chunara@nyu.edu
New York University

ABSTRACT

While machine learning is rapidly being developed and deployed in settings such as influenza prediction, there are critical challenges in using data from one environment to predict in another due to variability in features. Even within disease labels there can be differences (e.g. "fever" may mean something different reported in a doctor's office versus in an online app). Moreover, models are often built on passive, observational data which contain different distributions of population subgroups (e.g. men or women). Thus, there are two forms of instability between environments in this observational transport problem. We first conceptualize the underlying causal structure of this problem in a health outcome prediction task. Based on sources of stability in the model, we posit that we can combine environment and population information in a novel population-aware hierarchical Bayesian domain adaptation framework that harnesses multiple invariant components through population attributes when needed. We study the conditions under which invariant learning fails, leading to reliance on the environment-specific attributes. Experimental results for an influenza prediction task on four datasets gathered from different contexts show the model can improve prediction in the case of largely unlabelled target data from a new environment and different constituent population, by harnessing both environment and population invariant information. The proposed approach will have significant impact in many social settings wherein *who* the data comes from and *how* it was generated, matters.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

data generating process, domain adaptation, influenza prediction

ACM Reference Format:
Vishwali Mhasawade, Nabeel Abdur Rehman, and Rumi Chunara. 2020. Population-aware Hierarchical Bayesian Domain Adaptation via Multi-component Invariant Learning. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '20), April 2–4, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3368555.3384451

ACM CHIL '20, April 2–4, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7046-2/20/04…$15.00
https://doi.org/10.1145/3368555.3384451

1 INTRODUCTION

Machine learning algorithms have the potential to significantly improve prediction efforts across critically important healthcare tasks. Yet, there are several issues that must be addressed before the potential of machine learning in health is broadly realized. While individual models are built on and may perform well on a select dataset from a specific environment (also called "domain" in the literature) and population (e.g. the population could be skewed towards younger people or other demographics depending on where it's sampled from), improving prediction in new datasets gathered in different contexts and simultaneously, from different constituent populations, is a clear challenge articulated by many health practitioners [29].

First, standardization in health-related features is a significant problem. Variance in testing and billing practices [20, 24] as well as differences in clinical case definitions [25] from one environment to another present barriers for model transport. Accordingly, the same symptoms (features) can mean different things in different environments; "fever" may mean something different reported to a doctor at hospital *A* versus hospital *B*, or to a doctor compared to through a smartphone app [25, 26]. This issue is becoming more pertinent as the number and types of data collection environments (from clinical data, to healthworker-facilitated data wherein healthworkers visit individuals' houses, record symptoms and take specimens, to citizen-science studies in which participants report symptoms and submit specimens directly [13, 14]) is rapidly increasing. In all cases, obtaining labels can be impractical; e.g. for influenza they would require costly and time-consuming laboratory tests. Another critical challenge is that models are often built on data from a particular population in an environment, and transporting results to a different population can be challenging if subgroups are differently represented in source and target populations (representation bias [31]). These differences in data collection and demographic distributions make the problem of predicting infection in a dataset by using data gathered from different environments and populations challenging. We therefore address this unique problem of domain adaptation in the presence of representation bias. We study the problem via a simple, but important influenza prediction task.

The idea of transporting observational findings from source environment(s) to a target environment is essential in science and the concept has been well-defined on the basis that target environments can often differ from source environments. Furthermore, it can be expensive to generate labels in a new environment [22]. Methods have been proposed to exploit the causal structure of the data generating process in order to address certain domain adaptation problems, each relying on different assumptions. While some work has focused on identifying the invariant components to ensure robust transfer [17, 30], work by Pearl and Bareinboim [22] showed

that identifying the mechanisms by which two environments differ can also be used to inform empirical learning and incorporation of local variations in a system. With this background, in this paper we address the problem of observational transport with *both* environment differences and population representation bias. We do this by proposing a new hierarchical domain adaptation model that includes population attributes in the hierarchy in order to capture invariant information through these multiple components. The model then allows transfer of invariant information as well as learning information specific to a local environment *when necessary*. We are able to propose a solution to this problem by harnessing research in health regarding population structure (invariance in population attributes) along with algorithmic innovation to design this novel approach.

To accomplish this goal in a principled way, we first represent the data generating process (DGP) for our task via a selection diagram. Besides explicitly illustrating variables (nodes) and the mechanisms by which the nodes are assigned a value (edges) that are relevant to the DGP and do not vary across environments, a selection diagram includes S-variables which localize the mechanisms where sources of unreliability in the DGP exist. We formalize this description and discuss the selection diagram for the task in this study in following sections. We highlight that modeling the DGP requires an understanding of health concepts [22]. Thus for the task considered here (influenza prediction from symptoms) in order to identify the invariant and variant components of the causal graph, we leverage health research which shows that 1) reports of symptoms in relation to infection status vary by the data collection mode, and 2) while the population represented in an observational sample can suffer from selection bias, disease risk can be stratified by population groups [5, 28]. In societally-prescient problems such as health, attributes of whom the data is from (population demographics like age, gender) are commonly available, and it is understood that there are shared characteristics within these groups [28].

In sum, we specifically address a situation in which both environment and constituent population change from the source to target datasets; often the case in health prediction tasks. We use a simple but important task of influenza prediction from symptoms, and four real-world datasets representing a diverse set of environments and populations. Specific contributions are: 1) Formalizing the DGP between symptom reports and infection status, capturing sources of stability and of variance across environments (which we categorize into two: selection bias and feature instability); 2) A new domain/environment adaptation model for observational transport that accounts for instability in observed features as well as improves prediction on population subgroups even when not well represented in a particular dataset, through sharing invariant population characteristics in multiple components *as needed* (when a population subgroup is not well-represented in the target environment or its characteristic is different from that in other data); 3) Demonstrating the model on real-world data, showing significant improvement in prediction of infection on largely unlabelled target dataset overall *and* by population subgroups in comparison with several relevant baselines.

2 NOTATION AND PROBLEM SETTING

We consider source datasets from multiple environments $\mathcal{D}_e := \{(x_i^e, y_i^e, a_i^e, g_i^e)\}_{i=1}^{n_e}$ where $e \in E$ (E comprises of all the source environments) and a single target dataset

$$\mathcal{D}_t := \{(x_i^t, y_i^t, a_i^t, g_i^t)\}_{i=1}^{k} \bigcup \{(x_i^t, a_i^t, g_i^t)\}_{i=k+1}^{n_t}$$

where $k << n_t$; $t \in T$. For the target dataset we have limited number of labeled samples (k) whereas for the source datasets all the samples are labeled. L denotes all the datasets: source as well as the target ($L = E \cup T$). Sets of variables are denoted by italicized capital letters whereas lowercase letters are used for their individual assignments.

Y denotes the presence ($y = 1$) or absence ($y = 0$)) of the influenza virus. Age of the individual is represented using A, and categorized by common epidemiological groups: age 0-4, age 5-15, age 16-44, age 45-64, age 65+. Similarly, G represents gender (male or female). The demographic attributes (A and G, but can be expanded to other demographic attributes where possible) are together represented as D; $D = \{A, G\}$. X is the feature vector representing presence of the symptoms: fever, cough, muscle pain and sorethroat. Here x is a 4-dimensional binary vector representing the symptoms that an individual has (if an individual i has fever and sorethroat but no cough and muscle pain; the feature vector looks like $x_i = \{1, 0, 0, 1\}$). We consider subgroups in the data to be the specific demographic populations of interest belonging to a specific gender and age group $\mathcal{D}_{a,g} = \{(X, Y) \mid A = a, G = g\}$. The task is to predict the value of Y for each of the subgroups $\mathcal{D}_{a,g}$ from the symptom information X. This can be formalized as:

$$\min_{\forall a,g} R^t\left(f\left(X^t, \theta^t\right)\right) + \sum_e R^e\left(f\left(X^e, \theta^e\right)\right)$$

We aim to learn classifier $f\left(X^t, \theta^t\right)$ for the target dataset $\mathcal{D}_t$ parameterized by θ^t for each of the demographic subgroups ($\mathcal{D}_{a,g}$) that minimizes empirical risk R^t while minimizing total risk across the source environments R^e as well. It should be noted that the probability distribution of the target environment across population subgroups $P_t\left(X, Y \mid D\right)$ may not be uniform. Hence, the resulting $f\left(X^t, \theta^t\right)$ cannot be assumed to be the same across all subgroups.

3 RELATED WORK

Influenza Prediction Influenza is a global threat, affecting countries worldwide with considerable morbidity and mortality [27]. Globally, annual epidemics are estimated to result in about 3 to 5 million cases of severe illness, and about 290,000 to 650,000 respiratory deaths [36]. With the possibility of global pandemics looming, improving prediction of influenza is a continuing central priority of global health preparedness efforts. Predicting from symptoms in single datasets have used regression models [18], typically examining specific case definitions (sets of syndromic features). Machine learning approaches have enabled wider feature space examination [23]. While it is understood that health-related features can vary from hospital to hospital [35], influenza data sources incur even more diversity as passive observations are collected via varied sources including syndromic surveillance systems, Internet apps, and health worker based studies. Also, generating labels is difficult and costly (requires laboratory testing). Recent work has shown that domain adaptation can be useful for prediction from symptom

data sets obtained via these different environments [26]. While epidemiological study has indicated that there are disparities in risk by age group and gender for disease in general, and influenza specifically [1], prediction approaches that harness population attribute differences are an important gap in disease prediction models.

Observational transport. Observational transport refers to the transport of causal relationships across environments in which only passive observations can be collected [22]. The simple idea indicates that causal knowledge shows which mechanisms remain invariant under change. Accordingly, some work has used causal diagrams or feature selection methods to determine invariant relations in the source environment that can be transferred to the target environment, isolating the set of features which can be conditioned on to eliminate instabilities in the data generating process [17, 19, 30]. Though it should be noted that early work goes on to state that the causal relation to be transported can be learned from invariant components and *variant* components from both the source and target environments, depending on the DGP [22]. Here, we use this idea to allow trade-off between invariant characteristics across environments and empirical re-learning of relationships from each local environment, depending on which populations are represented in a dataset. In other words, we transmit invariant information through multiple population components, and use variant information *as necessary*, addressing the problem of different population subgroup representation in observational data.

Multi-source domain adaptation and hierarchical modeling. Domain adaptation is focused on improving performance for a target data set, in situations where the environment of the target data is different from the that of the source(s) from which information is transferred. Another approach to learning from multiple sources by pooling and analyzing multi-site datasets includes transforming the source and target feature spaces to correct any distributional shift in the data [37]. Prior work have also leveraged multiple source datasets to increase the amount of information learned [15]. This task has also been formulated from a causal view [19], where the posterior of the target is a weighted average of the source datasets. The "Frustratingly Easy Domain Adaptation" method is notable for simplicity and good performance on text data [8] and is equivalent to hierarchical domain adaptation [11] (except it explicitly ties parameters across environments). Hierarchical approaches, which have primarily been developed in natural language processing, in contrast allow hyperparameters to be separated across environments; each environment has its own environment-specific parameter for each feature which the model links via a hierarchical Bayesian global prior instead of a constant prior. This prior encourages features to have similar weights across environments unless there is good contrary evidence. This supports the goal of this work, to combine environment information as needed (unless the population represented in the local environment is much different than in other environments). Hierarchical Bayesian frameworks are a more principled approach for transfer learning, compared to approaches which learn parameters of each task/distribution independently and smooth parameters of tasks with more information towards coarser-grained ones [2]. In this work we advance this idea by creating a novel multi-level, multi-component hierarchy, as well as by the idea of incorporating population-attribute invariance as part of the hierarchy.

4 PROPOSED APPROACH

4.1 Assumptions

Here we describe the assumptions that ensure our problem is well-posed. The main assumption is that the data generating process is known and can be represented via a graphical causal diagram (helps to identify the information that can be transported [22]). We adapt the definition of a selection diagram which is previously defined [21, 22] to clearly delineate different types of change mechanisms.

Definition 1 (Selection diagram). *A selection diagram is a probabilistic causal model (as defined in [21]) augmented with auxiliary selection variables S (denoted by square nodes, which denote places of instability in the DGP) comprising of two types; $S = \{S^*, \tilde{S}\}$. An S^* variable can point to any observed variable. $S^* \rightarrow X$ denotes that the mechanism of assigning value to X changes across environments. The other type of selection variable $\tilde{S}$ represents a selection bias. Thus an edge from X to $\tilde{S}$ $\left(X \rightarrow \tilde{S}\right)$ denotes a non-random selection of individuals, groups or data for variable X.*

We can now formalize the causal and selection diagrams (Figure 1) for our setting (prediction of influenza infection from symptoms) based on prior knowledge and research in health. Along with the system variables: virus (Y), symptoms (X) and demographic attributes (D) of age and gender, we also have the selection variables $\left(S = \{S^*, \tilde{S}\}\right)$ which denote differences in the data-generating process across environments through instability in observed variables and selection bias. The symptoms that result are generally shaped by infection status [3], thus we have $Y \rightarrow X$. Population demographic attributes also can affect symptoms reported, X, and susceptibility to infection by the virus, Y (for example, symptoms common in young versus older people can vary; $X \leftarrow D, D \rightarrow Y \rightarrow X$) [5]. Now, we consider the parts of the data-generating process that vary across environments. The data collection environment (here, for example citizen science or health-worker facilitated) affects $P(X \mid Y)$ (specifically, it is known that symptoms reported via citizen science are less specific than in a hospital, for example) [25]. Thus the collection source introduces differences in the manner in which $P(X \mid Y)$ is observed across environments and there is a

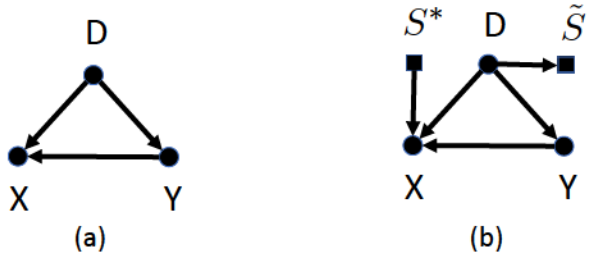


Figure 1: (a) Causal diagram, (b) Selection diagram representing the differences in the data-generating process.

selection variable pointing towards X; ($S^* \rightarrow X$). The absence of a selection variable pointing at D and Y indicates that the mechanism of assigning values to these variables is the same across environments (which makes sense intuitively, as demographic variables, e.g. man or woman, do not change or have different meanings in the different environments, nor does the process for obtaining flu infection status which is performed by laboratory confirmation in all cases). Finally, there is a selection bias associated with population demographic attributes. The proportion of individuals in each of the subgroups commonly varies across environments based on observational sampling (it is rare to have a representative distribution in a population sample unless an experiment is designed in advance and specific groups are recruited); $P_e(X, Y \mid D) \neq P_t(X, Y \mid D)$. Thus there is an edge from D to $\tilde{S}$. We now state the assumptions that help to formulate observational transport for this causal structure.

Assumption 1. *Let $\mathcal{G}$ be a causal graph with variables V consisting of the system variables $\mathcal{I} = \{X, Y, D\}$ and the selection variables $\mathcal{J} = \{S^*\}$.*

(1) *No system variable directly causes any selection variable ($\forall j \in \mathcal{J}, \forall i \in \mathcal{I} : i \rightarrow j \notin G$).*
(2) *No system variable is confounded by any selection variable ($S^*, \tilde{S}$).*

Assumption 2. *Let $\mathcal{G}$ be a causal graph with variables V consisting of the system variables $\mathcal{I} = \{X, Y, D\}$ and the selection variables $\mathcal{S} = \{S^*, \tilde{S}\}$ and $P(V)$ be the corresponding distribution on V.*

(1) *The distribution $P(V)$ is Markov and faithful with respect to $\mathcal{G}$.*
(2) S has no direct effect on Y ($S \rightarrow Y \notin G$))

4.2 Observational transport across environments

Motivated by the approach stated in [22] we aim to leverage a statistical relation, $R(P)$ to be learned from source environment(s) (characterized by probability distribution P) and transfer it to another (target) environment, $R(P^*)$, (characterized by probability distribution P^*) particularly when gaining complete information about that relationship in the target environment is costly. The definition of observational transportability in [22] (Definition 5), asserts that the relation to be transported has to be constructed from the source data as well as observations from the target data. As there is no control on the data-generating process (no intervention on any of the system variables, in contrast to experimental data) we cannot use *do*-calculus for formalizing the causal relation, and instead must use conditional independencies to understand the relationship between the outcome, Y and features X, by obtaining the joint probability distribution $P^*(X, Y, D)$. In the following section we identify invariant parts of this relation (which can be learned in combination with the source environment), and transferred as well as the target environment-specific relations (variant components) to be learned directly from the target dataset.

4.3 Multi-component invariant transfer

Having knowledge of the data-generating process via the graphical casual model $\mathcal{G}$, we identify the invariant conditional distributions that can be transferred from the source environment ($\mathcal{D}_e$) to the target environment ($\mathcal{D}_t$).

Indeed, according to the causal diagram in Figure 1b, we do not find a set of features (X) that d-separates S and Y, $S \not\perp\!\!\!\perp Y \mid X$. However, we do notice that $S \perp\!\!\!\perp Y \mid D$; the invariant information $P(Y|D)$ can be transferred across the environments. This follows from the fact the different demographic subgroups of the population share characteristics; for example, babies are known to be susceptible to certain infections as opposed to older people; strengthening the fact that the conditional distribution $P(Y|D)$ can be transferred across environments. However, we do need to learn $P^*(Y \mid X, D)$ for the target dataset since $S \not\perp\!\!\!\perp Y \mid X, D$. We therefore present an approach to learn the environment specific component ($P^*(Y \mid X)$)[1] as well as the population invariance $P(Y \mid D)$ from shared characteristics.

4.4 Formal framework of the undirected hierarchical multi-source Bayesian approach

Having identified the sources of variability and stability, we now can describe details of the model specific domain adaptation approach which enables learning $P^*(Y \mid X)$ and $P(Y \mid D)$, as described in the previous section. In the framework, the lowest level of the hierarchy represents the datasets (within each environment, in our case, citizen science or health-worker facilitated), $l \in L$, for each of which we have the labeled data $\mathcal{D}_l$ of the dataset l as shown in Figure 2. As in all Bayesian settings, the dataset parameters θ^l should represent the data $\mathcal{D}_l$ well. Here, θ^l are influenced by the environment-specific parameters (θ^c); θ^l are generated according to $P\left(\theta^l \mid \theta^c\right)$, where $c \in C$ is the collection mode and $\theta^c = \{\theta^{cs}, \theta^{hw}\}$ where θ^{cs} represents the parameters for the citizen-science collection mode and θ^{hw} represents the parameters for the health-worker supported collection mode. In the undirected hierarchical model we allow the environment specific parameters to have multiple parents and learn all parameters simultaneously. Accordingly, the environment parameters are generated according to the distribution $P(\theta^c \mid \theta^a, \theta^g)$. Here, we explicitly represent the population parameters; θ^a for $a \in A$, the different age group categories, and θ^g for genders $g \in G$, $\theta^d = \{\theta^a, \theta^g\}$ and $d \in D$. The model thus learns the invariant component parameters (θ^d) for the different demographic subgroups (ages 0-4, 5-15, 16-44, 45-64, 65+, males, females). Population parameters θ^a and θ^g have the root parameter θ^{pop} as the parent, which represents invariant information across all of the datasets, environments and population attributes, $P\left(\theta^{pop} \mid \theta^{par(pop)}\right) \equiv P(\theta^{pop})$. Then, the joint distribution is:

[1] $P^*(Y \mid X) = \sum_D P^*(X \mid Y, D) \cdot \frac{P(Y|D)}{P^*(X|D)}$

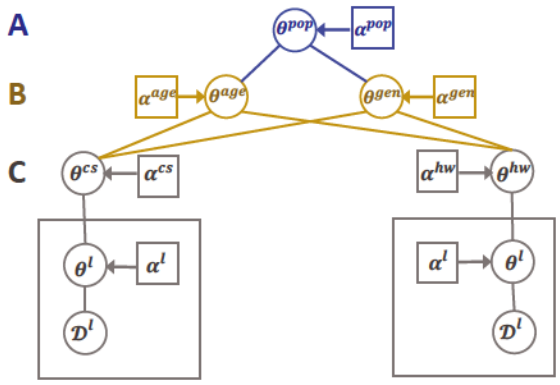


Figure 2: Population-aware hierarchical model; θ parameters at different nodes, $\mathcal{D}$ different data sets, α the priors. (A): Root level that represents invariant information across all data, (B): population parameters and information invariant to population-attributes (*age*) and (*gender*), (C): data set and environment-specific parameters and information (*cs* for citizen science and *hw* for healthworker facilitated datasets).

$$P(X,Y,\theta) = \prod_{l\in L} P\left(\mathcal{D}_l \mid \theta^l\right) \times \prod_{l\in L} P\left(\theta^l \mid \theta^c\right) \times \prod_{c\in C} P\left(\theta^c \mid \theta^a, \theta^g\right) \times \prod_{a\in A} P\left(\theta^a \mid \theta^{pop}\right) \times \prod_{g\in G} P\left(\theta^g \mid \theta^{pop}\right) \times P\left(\theta^{pop}\right)$$

We also study the conditions under which the invariant component parameters $\left(\theta^d\right)$ do not completely represent the information for a subgroup in which case the environment specific parameters $\left(\theta^l\right)$ help; thus explicating the conditions under which the invariant information is useful, and when environment-specific information should be utilized.

4.5 Hierarchy priors

For all parameters we use independent priors, computed based on symptom predictivity for each age group and gender. The inclusion of data dependent priors in Bayesian learning has been explored to incorporate domain knowledge into the posterior distribution of parameters [7]. For population-aware modeling, data-informed prior distributions are important because the distributions from each dataset are particular to the study, and thus capturing this information adds more information to the analysis than improper or vague priors (e.g. for a sample wherein one demographic group is under-represented), also motivates the multiple parents in the hierarchy. In contrast, using just the root prior for estimating the posterior ignores the demographic information available. Therefore, we use an empirical Bayes approach to specify weakly informative priors, centered around the estimates of the model parameters [34]. Root parameters are centered on the cumulative data since the root parameter captures environment invariant information.

4.6 Model steps

First, we use a probabilistic framework to jointly learn each parameter based on all levels of the hierarchy. We use a maximum a-posteriori parameter estimate instead of the full posterior for the joint distribution, which would be computationally intractable. We use a formulation, proposed in [9] that is amenable to standard optimization techniques, resulting in the objective:

$$F_{objective} = -\sum_{l\in L}\left[\sum_j (f_j + \lambda)\cdot\theta_j^l - \log\sum_k \exp(\theta_k^l)\right] + \beta \sum_{n\in Nodes} \mathrm{Div}(\theta^n, \theta^{par(n)}) \quad (1)$$

For dataset l, θ_j^l denotes the parameter for symptom j. From a specific dataset's parameter space, k represents individual symptoms. f_j is a statistical measure of the symptom j in the dataset, in this case the proportion of the particular symptom resulting in a positive influenza virus (i.e. the positive predictive value). *Nodes* is the set of all nodes in the hierarchy (here, $L \cup C \cup A \cup G$). Regularizing parameter λ was chosen as 1 to allow Laplacian smoothing. The function $\mathrm{Div}(\theta^n, \theta^{par(n)})$ is a divergence (L2 norm used) over the child and parent parameters that encourages child parameters (θ^n) to be influenced by parent parameters ($\theta^{par(n)}$), and allows a child parameter to be closely linked to more than one parent. The weight β represents the influence between node parameters and node parent parameters. Based on hyperparameter tuning, a value of 0.2 for β was used in all experiments. For objective function optimization we use Powell's method [12].

Second, we learn the influence (γ) of each parent on a particular dataset (child node). This is necessary since we need to learn $P^*(Y \mid X, D)$ for the target dataset as observed from the causal structure. We provide a mechanism to learn that as follows:

$$y_i^{(l,a,g)} = \gamma_0 + \gamma_1\left(\theta^l \cdot x_i^{(l,a,g)}\right) + \gamma_2\left(\theta^a \cdot x_i^{(l,a,g)}\right) + \gamma_3\left(\theta^g \cdot x_i^{(l,a,g)}\right)$$

The weights $\gamma_0, \gamma_1, \gamma_2, \gamma_3$ are estimated from a non-linear least square regression; the information from the different parents and the dataset can only be positive and hence we restrict the weights to be positive. This enables the model to give more weight to one level of the hierarchy when needed. In other words, how much demographic-invariant or environment-specific information is needed depends upon how much information is in a given dataset. For each of the subgroups a different classifier is learned based on the preferences of the subgroup. The reason for learning the weights for the different levels for each dataset independently is that each dataset would require different amounts of information from the demographic-specific and the environment-specific parameters, depending upon the demographic distribution of the sample in that dataset as well as the environment.

4.7 Licensing conditions for the use of invariant representations

To understand the cases under which the invariant representations captured by θ^a, θ^g fail to capture information for a specific subgroup, and local data must be used, we analyze information at the

demographic subgroup level. The model structure consists of different hierarchies wherein each hierarchical level learns invariant information. This implies that invariant information learned by the higher levels is invariant across environments as compared to the leaf nodes in which data-specific information is learned. We begin by describing the conditions on which information is evaluated.

Definition 3. *Let*
$P_{diff}(X \mid Y = y) = \left|P(X = 1 \mid Y = y) - P(X = 0 \mid Y = y)\right|$
be the difference of conditional probabilities of X (symptoms) given Y equal to y.

Definition 4. *Let*
$\delta_{\mathcal{D}} = \mathbb{E}_{x \in \mathcal{D}, y \in \mathcal{D}}\left[P_{diff}(X \mid Y = 1, A = a, G = g)\right]$
be the expectation of P_{diff} over the symptoms for the subgroup $\mathcal{D}_{a,g}$ of the dataset $\mathcal{D}$. Similarly we define δ_{pop} to be the expectation of P_{diff} over the symptoms for the population subgroup $\cup \mathcal{D}^{l}_{a,g}$ comprising of the subgroups from all the environments ($l \in L$).

Theorem 1. *The parameters θ for a subgroup ($\mathcal{D}_{a,g}$) of a dataset ($\mathcal{D}$) depends on the $\delta_{\mathcal{D}}$ and the conditional probability $P_{pop_{a,g}}(Y) = P(Y = 1 \mid l = pop, A = a, G = g)$ for the entire population comprising of the subgroups from the individual environments and the conditional probability $P_{\mathcal{D}_{a,g}}(Y) = P(Y = 1 \mid l = \mathcal{D}, A = a, G = g)$ for the subgroup of the specific dataset.*

$$\theta = \begin{cases} \theta^{l}, & \text{if } \delta_{\mathcal{D}} < \delta_{pop} \\ \theta^{l}, & \text{if } P_{\mathcal{D}_{a,g}}(Y) - P_{pop_{a,g}}(Y) \approx 1 \\ \theta^{d}, & \text{otherwise} \end{cases}$$

Proof Sketch. (full proof in Appendix)
a) We make use of the information function $I = -[\log(P_h)]$ which represents the information present about event h. If $\delta_{\mathcal{D}} < \delta_{pop}$ then $P(X = 1 \mid Y = 1, d = l) < P(X = 1 \mid Y = 1, d = pop)$ (this condition is explained in the proof in the appendix). Since I is a monotonically decreasing function, $I_d > I_{pop}$. Since the specific dataset has more information, the dataset specific parameters are used instead of using the invariant parameters learned over all the global population.
b) $P_{\mathcal{D}_{a,g}}(Y) - P_{pop_{a,g}}(Y) \approx 1$ if $P_{\mathcal{D}_{a,g}}(Y) \approx 1$ and $P_{pop_{a,g}}(Y) \approx 0$. This means that the specific subgroup (a, g) is over represented in the specific dataset l but we do not have much information about the specific subgroup from the invariant global representation since it is underrepresented in the global population. □

The conditions determine the cases in which spurious relations could be picked up by the invariant component representations θ^{d} and hence the data-specific parameters θ^{l} better represent the relations persistent in the specific dataset. The theorem states the conditions under which the invariant component representations θ^{d} will be used and when we need to rely on the data-specific parameters θ^{l} to capture the relations for a specific subgroup of the dataset.

5 DATA

Each dataset includes symptoms from individuals (X), laboratory confirmation of type of influenza virus they had (if any) (Y), and age and gender (D) of each person as example population attributes. Attributes of the datasets are summarized in Table 1, while the breakdown of positive and negative observations across demographics is shown in Figure 3. Through these differing study designs, one can see how the features may differ based on the stage of illness and included populations (e.g. those who have healthworkers come visit, versus those stay at home and may be sick but well enough to report on their own). As well, the difference in underlying populations illustrates the need for combining data in a principled way and accounting for these differences in the underlying sample composition. It should be emphasized that each of the datasets have a varied composition in terms of total number of observations and population demographics (Appendix Figure 1). We choose to use them all without any pre-processing, as these demonstrate real data set differences and will indicate model performance in such real-world situations. Indeed, population subgroups are not equally represented across all datasets. Goviral and Hongkong have the highest proportion of observations in the age group of 16-44, Fluwatch has the highest proportion of observations across the age group 45-64 while Hutterite has the highest proportion of observations in the age group of 5-15. The first two studies we classify as "citizen science" as they involve individuals self-reporting symptoms themselves from home, and taking their own nasal specimens for microbiological testing. The next two studies we classify as "healthcare worker" as they involve a trained worker visiting the individual, recording their symptoms according to a set criteria, and taking a nasal specimen from the participant. These studies also generally involved people more likely to be infectious [6]. Below we summarize the study design and context behind each dataset, including how the data was collected, while references are provided for the full papers describing all details.

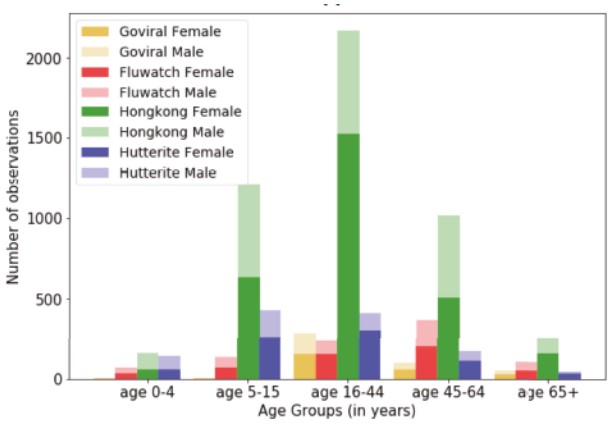


Figure 3: Demographic distributions in datasets. Darker shade of color denotes the number of females in the particular age group, lighter shade denotes number of males.

Table 1: Summary of dataset details.

Study	Location	Observations (positive)	Collection Type
Goviral	Northeast United States	520 (291)	Citizen Science
Fluwatch	England, United Kingdom	915 (567)	Citizen Science
Hongkong	Hong Kong	4954 (1471)	Healthworker Facilitated
Hutterite	Alberta, Canada	1281 (787)	Healthworker Facilitated

Table 2: AUC for flu prediction task (with 20% labeled data from target), bold values correspond to best performing model.

	Goviral	Fluwatch	Hongkong	Huttterite
TR	0.594	0.584	0.865	0.712
LR	0.585	0.490	0.914	0.706
FEDA	0.588	0.521	0.806	0.651
FEDA+pop	0.500	0.442	0.727	0.582
Hier	0.645	0.546	0.881	0.680
Hier+pop	**0.744**	**0.754**	**0.919**	**0.814**

The **GoViral** data comes from volunteers who self-reported symptoms online and also mailed in bio-specimens for laboratory confirmation of illness. These participants thus were never visited at home or in person at all. Volunteers were recruited, given a kit (collection materials and customized instructions), instructed to report their symptoms weekly, and when sick with cold or flu-like symptoms, requested to collect a nasal swab [14]. Periodic reminders were sent over email. Data from 2013–2017 is included.

FluWatch was a study in the United Kingdom, consisting of households which were recruited from registers of 146 volunteer general practices across England in seasonal and pandemic influenza over five successive cohorts from 2006–2011 [13]. Individuals participated from their home. In addition to a baseline visit by a nurse, households received participant packs containing paper illness diaries, thermometers and nasal swab kits including instructions on their use and the viral transport medium to be stored in the refrigerator. While participants would generate specimens and illness reports on their own, they were reminded every week via automated phone calls.

The **Hong Kong** study was focused on measuring infection in people who had household members who were already confirmed as sick. Household contacts of index patients (people who had come to the hospital and were confirmed to be sick) were followed in July and August 2009. These contacts live in close proximity with people who's illness was severe enough to take them to hospital, indicating a high risk for infection. Household members of 99 patients who tested positive for influenza A virus on rapid diagnostic testing were visited at their homes and swabs collected from household members by health workers over multiple weeks [6].

Data is also included from a study wherein nurses sampled people in **Hutterite** colonies in Alberta, Canada. The Hutterites are an ethno-religious group that tend to live together in colonies that are relatively isolated from towns and cities, and therefore are interesting places to examine respiratory infection prevalence given their self-contained nature. Data is from Dec. 2008 to June 2009 [16].

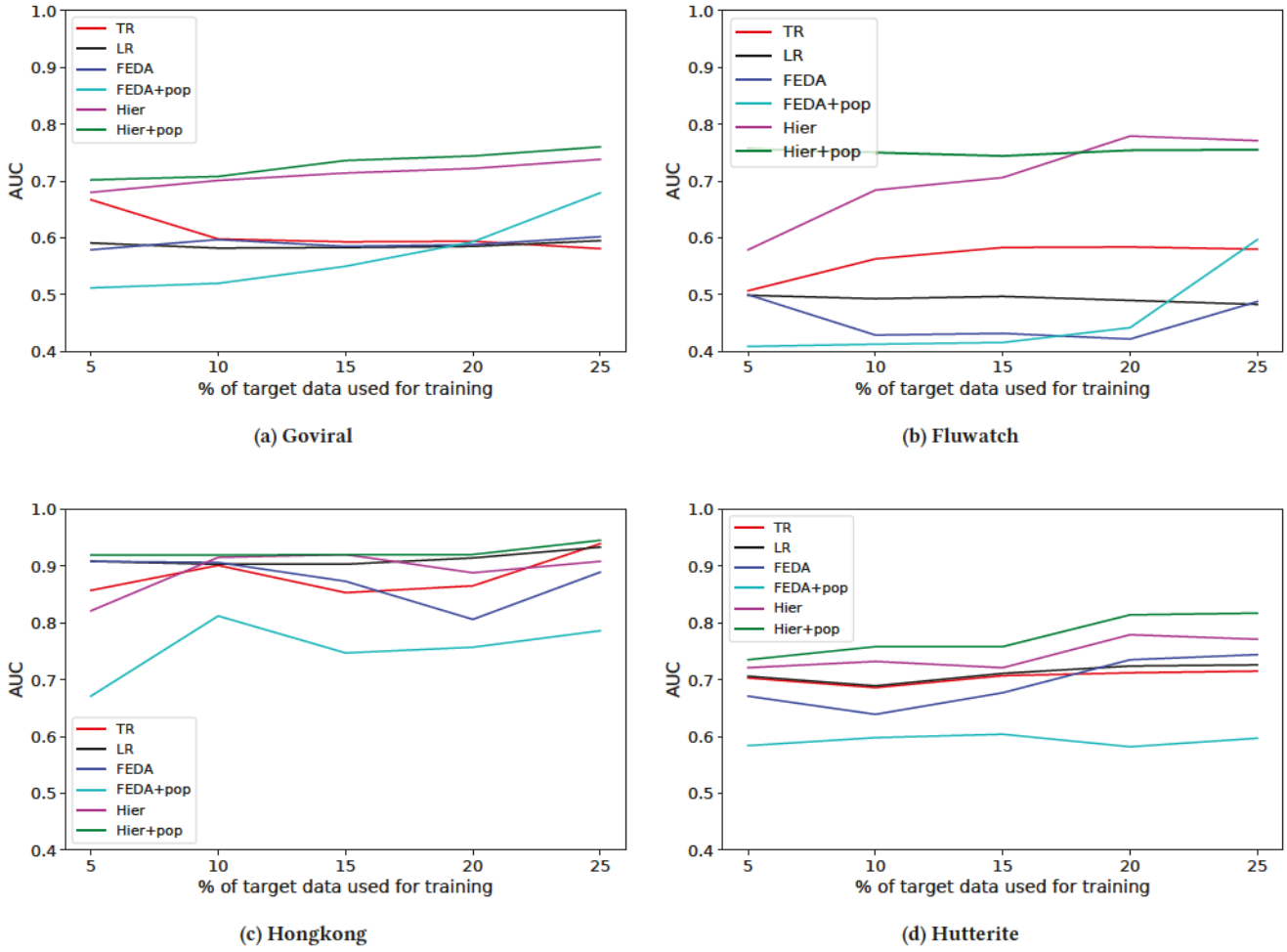
6 EXPERIMENTS

As motivated, we consider the case of transferring information from multiple source data sets from different domains to a largely unlabelled target dataset. We conduct multiple experiments to compare the proposed framework with relevant baselines to specifically examine the value of i) the hierarchical structure and ii) incorporation of population attributes, and iii) the amount of labelled data available from the target. Area under the ROC curve (AUC) metric is used to assess the performance based on both sensitivity and specificity. AUC is a measure of goodness of the ability of a binary classifier, equal to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. We evaluate AUC across all the population subgroups of the dataset ($D_{a,g}$). We compare results to three methods: Target only (**TR**), Logistic Regression (**LR**), Frustratingly Easy Domain Adaptation, which is noted for extreme simplicity and was used previously on symptom data [8, 26], without (**FEDA**) and with demographic attributes (**FEDA+pop**), Undirected Hierarchical Bayesian Domain adaptation without (**Hier**) and with demographic attributes (**Hier+pop**)[2]. Based on how the methods are designed, we describe how training and testing must work for each. The *Target only* method is trained only on a subset of the target environment dataset without incorporation of any information from source environments. *Logistic Regression* is trained on all the source datasets and a subset of the target dataset. *Frustratingly Easy Domain Adaptation* incorporates features specific to the source, specific to the target as well as a union of the source and target dataset, and is trained on both the source and subset of the target datasets. The proposed method, *Hier+pop*, is trained on both the source datasets and a subset of the target. All methods are tested on the target dataset (excluding the subset of the target data used for training).

7 RESULTS, DISCUSSION AND IMPACT

This work is relevant to the increasing scenarios in which algorithms are being used to co-analyse and use multiple real-world datasets, gathered from different contexts and composed of different population distributions. We present a novel approach in the framework of observational transport, applicable in scenarios

[2]Code is available at https://github.com/ChunaraLab/Pop-aware-domain-adaptation



(a) Goviral (b) Fluwatch

(c) Hongkong (d) Hutterite

Figure 4: Performance of Hier+pop method in comparison with baseline methods across increasing proportion of labelled target averaged across all population subgroups.

with instability in observed variables and selection bias; a significant challenge in many health transport problems. The model is motivated by knowledge of the underlying causal model. By testing on four real-world datasets for an influenza prediction from symptoms task, we show the multi-component model significantly improves performance by using principles of domain adaptation as well as by capturing information shared among population subgroups through a hierarchical and joint optimization approach. We perform a rigorous evaluation showing that with low amounts of labelled target data the model performs consistently better than baselines on entire datasets and on individual subgroups even when underrepresented in a specific dataset.

7.1 Performance analysis

Of the methods compared, TR and LR have the poorest performance (Table 2) across entire datasets. This makes sense, as a target-only model doesn't incorporate any information from other environments or populations. And, LR doesn't account for any population attributes. In all cases the Hier+pop method which accounts for the demographic attributes without including the demographic parameters explicitly in the same feature space as the symptoms (as is done by FEDA+pop), gives best performance across entire datasets. This also confirms the need to have different symptom parameters for specific demographic subgroups. We studied performance further based on amount of labelled training data available. We observe that Hier+pop performs consistently better than the baselines at low amounts of labelled target data (Figure 4). It should be noted that we examined results above 25% labels, and trends continue. As more labelled data becomes available, TR improves substantially as expected. Comparing datasets, Goviral has limited sample size (Figure 3), leading to low performance of baseline methods, and Hier+pop captures the invariant information across the source environments to improve performance over baselines drastically. As

Table 3: AUC scores across population subgroups († denotes use of θ^l otherwise θ^d is used) with 20% data used for training, bold values correspond to best performing model across population subgroup. Missing values (-) indicates the subgroup lacks data in either of the positive or negative classes, therefore models could not be trained for these subgroups.

Dataset	Method	Age 0-5		Age 5-15		Age 16-44		Age 45-64		Age 65+	
		Males	Females	Males	Females	Males	Females	Males	Females	Males	Females
Goviral	TR	-	-	-	-	0.655	0.796	0.661	0.524	0.238	0.692
	LR	-	-	-	-	0.541	0.845	0.774	0.607	0.188	0.542
	FEDA	-	-	-	-	0.638	0.845	0.728	0.464	0.112	**0.742**
	FEDA+pop	-	-	-	-	0.524	0.602	**0.807**	0.214	0.112	0.700
	Hier	-	-	-	-	0.681	0.767	0.766	0.645	0.312	0.700
	Hier+pop	-	-	-	-	**0.842**	**0.910**	**0.807**	**0.666**	**0.500†**	**0.742**
Fluwatch	TR	-	0.158	0.762	0.726	0.808	0.708	0.293	0.678	0.239	0.789
	LR	-	0.440	0.286	0.583	0.295	0.467	0.384	0.486	0.647	0.588
	FEDA	-	0.711	0.465	0.577	0.105	0.232	0.530	0.118	0.719	0.284
	FEDA+pop	-	0.440	0.298	0.565	0.311	0.412	0.389	0.260	0.803	0.505
	Hier	-	0.710	0.547	0.63	0.676	0.691	0.388	**0.767**	0.323	**0.833**
	Hier+pop	-	**0.868**	**0.747†**	**0.928†**	**0.787†**	**0.757†**	**0.757**	**0.767†**	**0.847**	**0.833†**
Hongkong	TR	-	0.156	0.961	0.963	0.954	0.959	0.997	0.878	0.929	**0.922**
	LR	-	0.156	0.957	0.960	0.948	0.960	0.997	0.880	1.000	**0.922**
	FEDA	-	0.156	0.936	0.943	0.921	0.873	0.957	0.810	0.864	0.797
	FEDA+pop	-	0.156	0.957	0.710	0.948	0.752	**0.997**	0.674	0.773	0.583
	Hier	-	0.211	0.975	0.990	0.991	0.993	0.980	0.930	**1.000**	**0.922**
	Hier+pop	-	**0.500†**	**0.995**	**1.000**	**0.995**	**0.996**	**0.997**	**0.939**	**1.000**	**0.922**
Hutterite	TR	0.714	0.750	0.780	0.995	0.358	0.860	0.741	0.971	0.286	0.320
	LR	0.532	0.902	0.904	0.793	0.431	0.780	0.76 9	0.783	0.405	0.180
	FEDA	0.532	0.902	0.904	0.793	0.431	0.780	0.769	0.783	0.405	0.180
	FEDA+pop	0.500	0.500	0.671	0.604	0.380	0.847	0.834	0.906	0.405	0.180
	Hier	0.831	**0.902**	**0.957**	0.792	0.380	0.760	0.834	0.760	0.404	0.180
	Hier+pop	**0.851**	**0.902†**	**0.957**	**1.000**	**0.576†**	**1.000**	**0.890**	**0.971**	**0.500**	**0.500**

compared to Goviral, Hutterite has better representation of the population subgroups and hence the baselines do not perform poorly but Hier+pop still performs substantially better. We highlight these results for Goviral and Hutterite datasets due to the vastly different sample sizes and data collection environment; results for other datasets follow the same trends (Figure 4). This demonstrates that multi-component invariant learning helps capture information shared among subgroups even when they are underrepresented. We also examined the learned parameters for the subgroups ($\mathcal{D}_{a,g}$), finding that they comply to conditions discussed in the Licensing conditions subsection. We also analyze performance of the methods by subgroup and find that Hier+pop indeed has better prediction across the subgroups, competing closely with TR in the case where θ^l are used instead of the invariant parameters θ^d (Table 3). In these specific cases, as expected, local information is preferred, therefore θ^l for the target dataset, which is influenced by the source environments, leads to a dip in the performance as compared to TR which does not have any influence by the source datasets.

7.2 Performance across population subgroups

Performance across subgroups for Goviral, Fluwatch, Hongkong and Hutterite is reported in Table 3. Hier+pop performs consistently better than the baselines across all the population subgroups for multiple datasets. We also report where the dataset specific parameters (θ^l) are used instead of the invariant (θ^d). This complies with the conditions provided in Theorem 1.

7.3 Performance across increasing proportion of labelled target data

Hier+pop performs consistently better than the baselines (TR, LR, FEDA, FEDA+pop) for multiple datasets. The trend is similar across the different datasets as shown in Figure 4. With increasing amounts of target data available, methods like Target only and FEDA perform well as more information about the dataset-specific feature space is available and there is less reliance on population-invariant information. In such scenarios the target dataset will have rich information about the target environment specific population subgroups.

7.4 Relations to fairness and social sciences

This work is motivated by real challenges health. We derive the proposed methodology in a principled manner from the causal structure of the problem. At the same time, there are overlaps with current research in fairness and the social sciences. First, formulation of the model in a hierarchical structure (versus with fixed effects) corroborates the latest thinking in social science theory, as

it doesn't treat characteristics across multiple types of parameters as additive [10]. Partitioning the variance across different aspects (symptoms and demographics), is meaningful as well, because the causes of variation at different levels may differ (e.g. causes of symptoms versus manifestations in different population subgroups). We derive this from the causal structure, but it also has been discussed in the epidemiology literature [32]. Moreover, hierarchical structure allows understanding of different datasets simultaneously rather than making reference to one perfect or representative dataset (e.g. by one reference group for each variable and parameter).

As well, while the aim of this study isn't to ensure specific fairness criteria are fulfilled mainly due to multiple sensitive attributes, the work still addresses an important problem identified in the fairness community: population subgroups can be represented differently in different datasets. Since the subgroups are characterized by multiple sensitive attributes (age and gender), some non-binary (age), imposing fairness constraints like demographic parity is not suitable especially when characteristics are shared across demographic subgroups. However, research in the fairness and machine learning literature has discussed how a one-size fits all model for all population subgroups suffers from aggregation bias when the conditional distribution $P(Y \mid X)$ is not consistent across the subgroups as shown by Suresh and Guttag [31]. We show here, for an influenza prediction task, that models that do not incorporate any subgroup-level information are not optimal across all the subgroups and are fitted to the dominant population subgroup. Hier+pop on the other hand mitigates the issue of aggregation bias by incorporating information across multiple specific subgroups.

One of the primary aims of formulating Hier+pop is thus to not compromise on sub-population AUC owing to selection bias and representation bias. We have drawn on principles such as beneficence ("do-the-best") and non-maleficence ("do-no-harm") and provide an approach to learn the parameters for each subgroup independently. As literature in fairness has discussed, model performance is often compromised due to inadequate sample sizes and appropriate data collection can aid in mitigating this issue [4]. Ustun et al. [33] have developed an approach to learn decoupled classifiers for each subgroup of the population, thereby ensuring fairness across the subgroups without performing any harm. Our work, capturing information in a hierarchical manner also is towards ensuring that performance on lesser sized subgroups is not comprised. While in large healthcare systems representation in multiple subgroups can be reached, often individual healthcare institutions or epidemiological studies, as demonstrated here, only reach a certain population distribution, and sourcing new data is simply not feasible. Accordingly, our approach by combining subgroup information across multiple environments is pragmatic. While methods like FEDA+pop and Hier performed at par with Hier+pop for a few select subgroups with adequately representative samples, Hier+pop performs consistently well across all the subgroups, and all four datasets.

7.5 Impact

As new datasets are constantly being generated in different environments and from different constituent populations, the model and findings from this work can be used in multiple ways by those designing surveillance systems. For example, to proactively assess and inform which population subgroups need to be further sampled to improve prediction in the target data (by comparing θ^l and θ^d across datasets). Knowledge of the criteria regarding local versus invariant parameters can also be used to identify which datasets can be combined to improve prediction. Practically, this work demonstrates how practitioners can save effort and cost by only labeling a proportion of data and combining data with other datasets to improve prediction. Overall, we present a practical approach to combine information from multiple instances especially when it is difficult to obtain labels; which can often be the case in public health, while retaining the global characteristics of population subgroups shared across environments.

8 ACKNOWLEDGMENTS

This work was supported in part by grant 1845487 from the National Science Foundation.

REFERENCES

[1] S. Bansal, B. Pourbohloul, N. Hupert, B. Grenfell, and L. A. Meyers. The shifting demographic landscape of pandemic influenza. *PLoS One*, 5(2):e9360, 2010.
[2] B. P. Carlin and T. A. Louis. *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall/CRC, 2010.
[3] CDC. Flu symptoms and diagnosis. https://www.cdc.gov/flu/symptoms/index.html, 2019.
[4] I. Chen, F. D. Johansson, and D. Sontag. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550, 2018.
[5] R. Chunara, E. Goldstein, O. Patterson-Lomba, and J. S. Brownstein. Estimating influenza attack rates in the united states using a participatory cohort. *Scientific reports*, 5:9540, 2015.
[6] B. J. Cowling, K. H. Chan, V. J. Fang, L. L. Lau, H. C. So, R. O. Fung, E. S. Ma, A. S. Kwong, C.-W. Chan, W. W. Tsui, et al. Comparative epidemiology of pandemic and seasonal influenza a in households. *NEJM*, 362(23):2175–2184, 2010.
[7] W. F. Darnieder. *Bayesian methods for data-dependent priors*. PhD thesis, The Ohio State University, 2011.
[8] H. Daume III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.
[9] G. Elidan, B. Packer, G. Heitz, and D. Koller. Convex point estimation using undirected bayesian transfer hierarchies. *arXiv preprint arXiv:1206.3252*, 2012.
[10] C. R. Evans, D. R. Williams, J.-P. Onnela, and S. Subramanian. A multilevel approach to modeling health inequalities at the intersection of multiple social identities. *Social Science & Medicine*, 203:64–73, 2018.
[11] J. R. Finkel and C. D. Manning. Hierarchical bayesian domain adaptation. In *NAACL HLT*, pages 602–610, 2009.
[12] R. Fletcher and M. J. Powell. A rapidly convergent descent method for minimization. *The computer journal*, 6(2):163–168, 1963.
[13] E. B. Fragaszy, C. Warren-Gash, L. Wang, A. Copas, O. Dukes, W. J. Edmunds, N. Goonetilleke, G. Harvey, A. M. Johnson, J. Kovar, et al. Cohort profile: The flu watch study. *International journal of epidemiology*, 46(2):e18–e18, 2016.
[14] J. Goff, A. Rowe, J. S. Brownstein, and R. Chunara. Surveillance of acute respiratory infections using community-submitted symptoms and specimens for molecular diagnostic testing. *PLoS currents*, 7, 2015.
[15] J. Guo, D. J. Shah, and R. Barzilay. Multi-source domain adaptation with mixture of experts. *arXiv preprint arXiv:1809.02256*, 2018.
[16] M. Loeb, M. L. Russell, L. Moss, K. Fonseca, J. Fox, D. J. Earn, F. Aoki, G. Horsman, P. Van Caeseele, K. Chokani, et al. Effect of influenza vaccination of children on infection rates in hutterite communities: a randomized trial. *JAMA*, 303(10):943–950, 2010.
[17] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10869–10879, 2018.
[18] A. S. Monto, S. Gravenstein, M. Elliott, M. Colopy, and J. Schweinle. Clinical signs and symptoms predicting influenza infection. *Archives of internal medicine*, 160(21):3243–3247, 2000.
[19] J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*, 2016.
[20] S. Mullainathan and Z. Obermeyer. Who is tested for heart attack and who should be: Predicting patient risk and physician error. Technical report, National Bureau of Economic Research, 2019.

[21] J. Pearl. Causality: models, reasoning, and inference. *IIE Transactions*, 34(6): 583–589, 2002.
[22] J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.
[23] A. L. Pineda, Y. Ye, S. Visweswaran, G. F. Cooper, M. M. Wagner, and F. R. Tsui. Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *JBI*, 58:60–69, 2015.
[24] R. Pivovarov, D. J. Albers, G. Hripcsak, J. L. Sepulveda, and N. Elhadad. Temporal trends of hemoglobin a1c testing. *JAMIA*, 21(6):1038–1044, 2014.
[25] B. Ray and R. Chunara. Predicting acute respiratory infections from participatory data. *Online journal of public health informatics*, 9(1), 2017.
[26] N. Rehman, M. M. Aliapoulios, D. Umarwani, and R. Chunara. Domain adaptation for infection prediction from symptoms based on data from different study designs and contexts. *arXiv preprint arXiv:1806.08835*, 2018.
[27] N. G. Reich, L. C. Brooks, S. J. Fox, S. Kandula, C. J. McGowan, E. Moore, D. Osthus, E. L. Ray, A. Tushar, T. K. Yamana, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, 2019.
[28] S. Saria, D. Koller, and A. Penn. Learning individual and population level traits from clinical temporal data. In *Proceedings of Neural Information Processing Systems*, 2010.
[29] H. Singh, R. Singh, V. Mhasawade, and R. Chunara. Fair predictors under distribution shift. *arXiv preprint arXiv:1911.00677*, 2019.
[30] A. Subbaswamy, P. Schulam, and S. Saria. Learning predictive models that transport. *arXiv:1812.04597*, 2018.
[31] H. Suresh and J. V. Guttag. A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002, 2019. URL http://arxiv.org/abs/1901.10002.
[32] M. Susser and E. Susser. Choosing a future for epidemiology: Ii. from black box to chinese boxes and eco-epidemiology. *American journal of public health*, 86(5): 674–677, 1996.
[33] B. Ustun, Y. Liu, and D. Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382, 2019.
[34] S. van Erp, J. Mulder, and D. L. Oberski. Prior sensitivity analysis in default bayesian structural equation modeling. *American Psychological Association*, 2017.
[35] J. Wiens, J. Guttag, and E. Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *JAMIA*, 21(4): 699–706, 2014.
[36] World Health Organization. Influenza fact sheet: Overview. 2018.
[37] H. H. Zhou, V. Singh, S. C. Johnson, G. Wahba, A. D. N. Initiative, et al. Statistical tests and identifiability conditions for pooling and analyzing multisite datasets. *PNAS*, 115(7):1481–1486, 2018.