Bayesian Nested Latent Class Models for Cause-of-Death Assignment using Verbal Autopsies Across Multiple Domains

Zehang Richard Li\textsuperscript{1}, Zhenke Wu\textsuperscript{2}, Irena Chen\textsuperscript{2}, and Samuel J. Clark\textsuperscript{3}

\textsuperscript{1}Department of Statistics, University of California, Santa Cruz
\textsuperscript{2}Department of Biostatistics, University of Michigan
\textsuperscript{3}Department of Sociology, The Ohio State University

June 26, 2023

Abstract

Understanding cause-specific mortality rates is crucial for monitoring population health and designing public health interventions. Worldwide, two-thirds of deaths do not have a cause assigned. Verbal autopsy (VA) is a well-established tool to collect information describing deaths outside of hospitals by conducting surveys to caregivers of a deceased person. It is routinely implemented in many low- and middle-income countries. Statistical algorithms to assign cause of death using VAs are typically vulnerable to the distribution shift between the data used to train the model and the target population. This presents a major challenge for analyzing VAs as labeled data are usually unavailable in the target population. This article proposes a Latent Class model framework for VA data (LCVA) that jointly models VAs collected over multiple heterogeneous domains, assigns causes of death for out-of-domain observations, and estimates cause-specific mortality fractions for a new domain. We introduce a parsimonious representation of the joint distribution of the collected symptoms using nested latent class models and develop a computationally efficient algorithm for posterior inference. We demonstrate that LCVA outperforms existing methods in predictive performance and scalability. Supplementary materials for this article and the R package to implement the model are available online.

Keywords: Domain adaptation; Data shift; Classification; Mixture model; Dependent binary data; Quantification learning.

ZRL was supported by grant R03HD110962 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). ZW and IC were supported in part by a seed grant from Michigan Institute of Data Science. SJC was supported by grant R01HD086227 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD).
# 1 Introduction

Data describing cause of death is an essential component for understanding the burden of disease, emerging health needs, and effect of public health interventions. Few low- and middle-income countries (LMIC) have systems that produce high quality cause of death statistics. Only about two-thirds of the deaths worldwide are registered and up to half of these deaths are either not assigned a cause or assigned only an ill-defined cause (World Health Organization [2021]). In many populations not served by official medical certification of causes, a technique known as verbal autopsy (VA) has been routinely used to infer causes of death. VA is a survey-based method whereby a structured questionnaire is conducted to a family member or a caregiver of a recently deceased person after a suitable mourning period. The VA interview collects information about the circumstances, signs, and symptoms leading up to the death. It is routinely used by researchers in Health and Demographic Surveillance System (HDSS) including the INDEPTH network (Sankoh and Byass [2012]) and the ALPHA network (Maher et al. [2010]), multi-country research projects (Nkengasong et al. [2020]; Breiman et al. [2021]), and national scale surveys in many LMICs.

Data collected by VAs are analyzed either by a panel of physicians or with statistical algorithms. The physician review approach can be effective if resource permits (Lozano et al. [2011]), but it is expensive and rarely possible for timely monitoring of cause-specific mortality. On the other hand, statistical methods have been widely used to automate the cause-of-death assignment from VA data (e.g., Byass et al. [2003]; Miasnikof et al. [2015]; Serina et al. [2015]; McCormick et al. [2016]). These methods aim to construct a high-dimensional classifier for cause-of-death assignment and obtain population-level cause-specific mortality fraction (CSMF) estimates, i.e., the distribution of deaths due to each possible cause.
There are many methodological challenges for cause-of-death assignment algorithms to be successfully deployed in practice. First, all existing VA algorithms assume that there is a single training dataset or domain knowledge base from which the relationship between symptoms and causes can be learned. In most situations where VA is used, local deaths from the target population with known causes are extremely rare. High-quality VA data with reference causes are usually only available outside of the target population, and may consist of data collected from several distinct study populations. This leads to the data shift problem, where the source and target domain data may have different marginal and conditional distributions (Moreno-Torres et al., 2012), and it could significantly bias the cause-of-death assignment and CSMF quantification. The issue of data shift received little discussion in the midst of developments of automated VA cause-of-death assignment algorithms in the last decade. The recent work of Datta et al. (2021) and Fiksel et al. (2021) are the first to address this issue by calibrating model predictions to additional local validation data with known causes of death. When no validation data exist, however, all existing VA methods are vulnerable to data shift. To our knowledge, there is no literature that characterizes the heterogeneous joint distributions of symptoms and causes across different domains or leverages such information for cause-of-death assignment.

Second, existing VA methods used by practitioners typically assume that symptoms are conditionally independent given the underlying cause of death as it reduces model complexity and computation cost significantly (Byass et al., 2003; Miasnikof et al., 2015; McCormick et al., 2016). More recently, several approaches have been proposed to account for symptom dependence using latent Gaussian models with sparse or low-rank representations (Li et al., 2020; Kunihama et al., 2020; Moran et al., 2021). It has been shown that incorporating symptom dependence usually improves the accuracy of both individ-
ual cause-of-death assignment and the population-level CSMF estimation. However, the inferred dependence relationship among the continuous latent variables is difficult to interpret, as they do not correspond to the dependence of the observed binary symptoms (Li et al., 2020). In addition, the computation required to estimate the high-dimensional latent Gaussian mixtures can take hours or days to complete, which makes these methods difficult for routine use in low-resource settings.

In this paper, we address these challenges by developing a novel Latent Class model framework for VA data (LCVA). A key feature of LCVA is that we estimate a common collection of latent symptom profiles across different populations. This leads to an explicit domain adaptation strategy for classifying out-of-domain deaths that takes into account the similarity between the target population and the existing training datasets. We introduce a sparse parameterization to further reduce the dimensionality of the latent symptom profiles that also facilitates direct interpretations of the conditional independence relationship among the observed symptoms. Furthermore, our inference procedure is an order of magnitude faster than the existing methods using latent Gaussian models to capture symptom dependence.

The rest of the paper is organized as follows. Section 2 reviews the background of existing VA methods and their implications under data shift. Section 3 proposes the nested latent class model approach for cause-of-death assignment. Section 3.1 discusses the model with one training dataset. Section 3.2 extends the model to the scenario where training data consist of deaths from multiple domains. Section 4 develops an efficient Markov chain Monte Carlo (MCMC) algorithm for posterior sampling. Section 5 demonstrates the proposed model under various realistic scenarios using the Population Health Metrics Research Consortium (PHMRC) gold-standard VA dataset. Section 6 concludes the paper.
and discusses future directions.

2 Domain adaption in cause-of-death assignment

The domain adaptation problem refers to the situation where a statistical learning model trained on one labeled dataset needs to be generalized to the target dataset, or target domain, drawn from a different distribution and with insufficient labeled data (Daume III and Marcu, 2006). In the context of cause-of-death assignment using VA, the discussion of cross-domain generalizability goes back to the early work of King and Lu (2008), where the authors developed a constrained least square model trained on hospital deaths with known causes and subsequently applied to classifying community deaths. Let $X_i$ denote the vector of symptoms for individual $i$ and $Y_i$ denote the corresponding cause of death. For simplicity, let us assume for now that there are two datasets: a training dataset $\mathcal{T}$ with both $X_i$ and $Y_i$ known, and a target dataset $\mathcal{T}_0$ where only $X_i$ is observed. King and Lu (2008) makes the assumption that $p_{T_0}(X|Y) = p_{T}(X|Y)$. This assumption implies that the difference in the joint distribution of $p(X,Y)$ between the two domains can be fully explained by the difference in $p(Y)$, since $p(X,Y) = p(Y)p(X|Y)$. Prediction tasks under this decomposition are also termed anticausal prediction (Schölkopf et al., 2012) and this type of data shift with domain-specific $p(Y)$ and domain-invariant $p(X|Y)$ is known as label shift or prior shift (Storkey, 2009).

A large collection of VA algorithms, however, took a more generic classification perspective. They first learn a model $p_{\mathcal{T}}(Y|X)$ with the training data and then apply the classifier to observations from the target domain for cause-of-death assignment. The generalizability of this approach relies on the assumption that $p_{T_0}(Y|X) = p_{\mathcal{T}}(Y|X)$. That is, the difference in the joint distribution of $p(X,Y)$ across domains are fully explained by
different symptom distribution $p(X)$. This type of data shift is also known as *covariate shift* (Shimodaira, 2000). Under this assumption, the target CSMF can be estimated by simple aggregation of individual classifications, i.e.,

$$
\hat{p}_T(Y) = \int_x \hat{p}_{T_0}(Y|X = x)p_{T_0}(x)dx = \int_x \hat{p}_T(Y|X = x)p_{T_0}(x)dx \approx \frac{1}{n} \sum_{i=1}^{n} \hat{p}_T(Y|X_i).
$$

Algorithms that make the implicit assumption on covariate shift include some of the most widely adopted VA models, such as InterVA (Byass et al., 2012, 2019), Tariff (Serina et al., 2015), Naive Bayes Classifier (Miasnikof et al., 2015), as well as many others in the literature applying off-the-shelf classification models on VA data (e.g., Flaxman et al., 2011, Blanco et al., 2020). Models assuming covariate shift are easy to design and implement, but difficult to justify in the context of VA. As pointed out by King and Lu (2008), symptoms collected by VA are usually the consequences of the underlying causes of death. Thus the covariate shift assumption is almost always violated as $p_{T_0}(Y|X)$ and $p_T(Y|X)$ are different even when $p(X|Y)$ is the same across domains.

The recent advancement of Bayesian modeling of VA data again returned to the formulation using the more natural decomposition of $p(X,Y) = p(Y)p(X|Y)$ to describe the data generating process. The first Bayesian hierarchical model for cause-of-death assignment, InSilicoVA (McCormick et al., 2016), proposed to use non-informative priors to model $p_{T_0}(Y)$ in the target population and strong informative priors to shrink $p_{T_0}(X|Y)$ to the observed $p_T(X|Y)$ calculated from training data or provided by domain experts. This framework was later extended to include more flexible characterizations of $p(X|Y)$, using latent Gaussian graphical models (Li et al., 2020) and factor models (Kunihama et al., 2020, Moran et al., 2021). While Kunihama et al. (2020) and Moran et al. (2021) treated the cause of death of all observations as drawing from the same distribution, they can be readily adapted to have different $p(Y)$ in different populations.
The prior shift assumption, however, can still be overly simplistic in practice. Figure 1 illustrates an example from the PHMRC gold-standard VA dataset (Murray et al., 2011a) discussed in Section 5. It shows the proportion of deaths with a ‘yes’ response in two symptoms, having trouble breathing and drinking alcohol, conditional on the ten causes of death with most observations, for each of the six study sites in the dataset. While it only illustrates the conditional distribution of two symptoms, strong domain heterogeneity can be observed from these site-specific empirical response probabilities \( p(X|Y) \). Some of these reporting patterns may be explained by the presence of confounding factors (Moran et al., 2021). However, we usually do not have enough data to identify or adjust for all potential confounders, many of which may be unobservable, or quantify the similarity of these distributions without reference cause of death. When \( p(X|Y) \) vary across domains, the choice of training data could have strong implications for out-of-domain predictions.

A line of research closely related to this paper is the calibration models developed in Datta et al. (2021) and Fiksel et al. (2021). The focus of these methods are fundamentally different from our work. They consider a different scenario where the cause of death is known for a small sample of observations \( T_0' \) from the target population, and assume that \( p_{T_0'}(X|Y) = p_{T_0}(X|Y) \), i.e., only prior shift may exist between the labeled and unlabeled deaths from the same population, with potentially different \( p_{T_0'}(Y) \) and \( p_{T_0}(Y) \). It can be shown that this assumption is sufficient to de-bias a classifier trained on \( T \) and applied to \( T_0 \), using the labeled target population data \( T_0' \). The advantage of such calibration methods is that they circumvent the need to identify and parameterize transportable component of the joint distribution from the training domains to the target domain. The cost, however, is the requirement for additional local data with known causes, which is usually impractical in resource-constrained settings. Moreover, the estimation of the error transition matrix
Figure 1: Empirical proportion of ‘yes’ response in two symptoms, drinking alcohol (left), and having trouble breathing (right), among deaths due to different causes of death in the six study sites in the PHMRC gold-standard dataset. Gray color indicates no collected death due to a given cause in a given site.

of the predicted cause of death is challenging when there are many potential causes of death. For a small number of local labeled data, Datta et al. (2021) and Fiksel et al. (2021) recommend grouping the causes of death into a handful of broad categories in the analysis. It remains an open problem to estimate the CSMF for longer cause lists with limited calibration data.

The focus of this paper is to develop a general framework to characterize domain heterogeneity and leverage such information when assigning causes of death for out-of-domain observations, without additional calibration data. Unlike any existing VA models, we allow both \( p(Y) \) and \( p(X|Y) \) to vary across domains. The key to our modeling approach is a latent class model representation (Goodman, 1974) that assumes source component shift for data across different domains (Storkey, 2009). That is, we assume that there is a latent
variable $Z$ that renders $p_T(X|Y, Z) = p_{T_0}(X|Y, Z)$. This assumption can be considered as an extension of prior shift on the latent space. It allows us to combine data from multiple domains with domain-specific $p(X, Y)$ efficiently when assigning cause of death for unlabeled deaths in a new domain. We discuss our proposed model in more details in the next section.

3 Nested latent class models for VA data

In this section, we present the nested latent class model for VA data, LCVA. We first introduce a single-domain latent class model that parsimoniously models the joint distribution of symptoms and causes, and then extend it to modeling data from multiple domains.

3.1 The single-domain model

Let $Y_i \in \{1, ..., C\}$ denote the categorical cause of death for the $i$-th death and $X_i \in \{0, 1\}^p$ denote the $p$-dimensional reported binary symptom vector. The standard practice of VA usually involves pre-processing any continuous variables into binary indicators (Li et al., 2020). Thus we focus our discussion on binary symptoms here, but the model can easily incorporate categorical symptoms of more than two levels, thus mitigating the potential loss of information in the pre-processing step. We begin by assuming that each cause of death can be further divided into $K$ sub-categories and each individual is associated with a latent class $Z_i \in \{1, ..., K\}$. We assume reported symptoms are independent draws from a Bernoulli distribution given both the cause of death $Y_i$ and the latent class membership $Z_i$, i.e.,

$$X_{ij}|Y_i = c, Z_i = k \sim \text{Bern}(\theta_{ck}), \quad j = 1, ..., p.$$
We treat the cause of death $Y = (Y_1, ..., Y_n)$ and latent class indicator $Z = (Z_1, ..., Z_n)$ as random variables and let, for $i = 1, ..., n$,

$$
Y_i \sim \text{Cat}(\pi),
$$

$$
Z_i | Y_i = c \sim \text{Cat}(\lambda_c).
$$

For deaths due to each cause of death, $p(X_i | Y_i = c) = \sum_{k=1}^{K} \lambda_{ck} \prod_{j=1}^{p} \theta_{ckj}^{x_{ij}} (1 - \theta_{ckj})^{1-x_{ij}}$. Conditional dependence among symptoms are introduced by marginalizing out the latent class indicators. With sufficiently many latent classes, this representation is flexible enough to represent any multivariate discrete distribution (Dunson and Xing, 2009). We treat the latent class $Z$ as defined within each cause of death and both latent indicators $Z$ and $Y$ need to be estimated for a new death. This nested formulation allows the model to be sufficiently expressive for deaths from each cause.

The naive factorization of the joint distribution of $(X, Y, Z)$ requires $CK(p + 1) - 1$ parameters. This can be computationally challenging for VA data, as the number of symptoms $p$ and the size of cause list $C$ are usually large while sample size is typically small to moderate. We adopt a sparsity-inducing prior to model the response probabilities $\theta_{ckj}$ proposed by Zhou et al. (2015). We let

$$
\theta_{ckj} = \delta_{ckj} \phi_{ckj} + (1 - \delta_{ckj}) \gamma_{cj}, \quad \delta_{ckj} \sim \text{Bern}(\tau_c),
$$

$$
\phi_{ckj} \sim \text{Beta}(1, \nu_\phi), \quad \gamma_{cj} \sim \text{Beta}(a_\gamma, b_\gamma), \quad \tau_c \sim \text{Beta}(1, \nu_\tau)
$$

where $\delta_{ckj}$ is a binary indicator specifying whether $p(X_{ij} | Y_i = c, Z_i = k)$ is affected by the latent class membership or takes the baseline response probability $\gamma_{cj}$. For each cause $c$, this prior encourages the associated latent classes $k = 1, ..., K$ to have similar response profiles that differ only in a subset of symptoms $\{j, \delta_{ckj} \neq 0\}$. Thus it significantly improves the efficiency of the latent representation when $K$ is large. In the original proposal of Zhou et al. (2015),
the baseline response vector is set to fixed values in advance. Here we model the baseline response probabilities \( \gamma_{cj} \) in a data adaptive fashion, as the baseline distribution of \( p(X_{ij}|Y_i = c) \) is generally unknown and hard to specify. In certain situations, we may have external information on such conditional distributions in the form of physician provided domain knowledge (McCormick et al. 2016; Li et al. 2020) which may be used to construct informative priors on \( \gamma_{cj} \). Here we use non-informative priors with \( \nu_\phi = a_\gamma = b_\gamma = 1 \).

We also expect deaths due to different causes to exhibit varying levels of complexity in the reported symptoms. For example, distribution of symptoms among deaths from external causes (e.g., drowning) might require fewer parameters to characterize than those from deaths due to infectious diseases. Instead of specifying different latent class size \( K \) for each cause, the complexity of the latent symptom profiles are governed by the cause-specific sparsity level \( \tau_c \). We use a non-informative prior on \( \tau_c \) with \( \nu_\tau = 1 \). We note that the sparsity induced by the mixture prior on \( \theta \) does not perform variable selection, but rather it aims to reduce the number of parameters by identifying symptoms that are conditionally independent of others. When \( \{\delta_{c1j},...,\delta_{cKj}\} \) are all 0’s for symptom \( j \) and cause \( c \), it implies that symptom \( j \) is independent of all other symptoms given cause \( c \). The conditional independence model assumed by the InSilicoVA algorithm can be thought of as the special case where \( \delta = 0 \).

To complete the hierarchical specification with priors for the latent parameters, we consider the stick-breaking prior for the mixing weights \( \lambda \) and a Dirichlet prior on \( \pi \), i.e.,

\[
\lambda_{ck} = V_{ck} \prod_{l<k}(1 - V_{cl}), \quad V_{ck} \sim \text{Beta}(1, \omega_c), \\
\omega_c \sim \text{Gamma}(a_\omega, b_\omega), \quad \pi \sim \text{Dirichlet}(\alpha).
\]

With a training dataset \( \mathcal{T} = \{\mathbf{x}_1, y_1, ..., \mathbf{x}_n, y_n\} \), the posterior predictive distribution for the cause of death of a single new data point \( (X^{(0)}, Y^{(0)}) \) from the same population can
be readily computed by
\[
p(Y^{(0)} = c|X^{(0)}, T) = \int p(Y^{(0)} = c|X^{(0)}, \pi, \lambda, \theta)p(\pi, \lambda, \theta|T)d(\pi, \lambda, \theta).
\]

VAs are most commonly used to determine the CSMF of a target population. Consider a target dataset \( T_0 = \{x_1^{(0)}, \ldots, x_n^{(0)}\} \) with unknown cause of death, we assume the same data generating process with target-specific distributions of causes of death and latent classes. For \( i = 1, \ldots, n \), we let
\[
Y_i^{(0)} \sim \text{Cat}(\pi_i^{(0)}),
\]
\[
Z_i^{(0)}|Y_i^{(0)} = c \sim \text{Cat}(\lambda_c^{(0)}),
\]
\[
X_{ij}^{(0)}|Y_i^{(0)} = c, Z_i^{(0)} = k \sim \text{Bern}(\theta_{ckj}).
\]

For the prior on the target CSMF, we let \( \pi^{(0)} \sim \text{Dirichlet}(\alpha) \). We consider two different specifications for \( p_{T_0}(X|Y) \):

1. Constant weights: The most restrictive parameterization is to assume only prior shift exists across the domains. That is, we let \( p_{T_0}(X|Y) = p_T(X|Y) \), or equivalently, \( \lambda^{(0)} = \lambda \).

2. New weights: We may relax the prior shift assumption by allowing the target dataset to have a different composition over the latent classes learned in the training domain. In other words, we assume \( p_{T_0}(X|Y, Z) = p_T(X|Y, Z) \) only. The new mixing weights can be estimated with the same prior as in the first stage, i.e.,
\[
\lambda^{(0)} \sim V_{ck}^{(0)} \prod_{l < k}(1 - V_{cl}^{(0)}), \quad V_{ck}^{(0)} \sim \text{Beta}(1, \omega_c^{(0)}), \quad \omega_c^{(0)} \sim \text{Gamma}(a_\omega, b_\omega)
\]

We note that the additional flexibility of the second predictive strategy may sometimes suffer from overfitting in the target domain when \( K \) is large. With data from multiple domains, we could further estimate \( \lambda^{(0)} \) in a more robust fashion, as described in the next subsection.
3.2 The multi-domain model

When training data are collected from a single domain, only one set of mixing weights can be inferred. Thus the mixing weights in the target domain are treated either as the same as those from the training domain, or as independent random variables. When the training data consist of observations from multiple domains, however, we can leverage the estimated collection of mixing weights to improve out-of-domain prediction. Consider a training dataset consisting of observations from $G$ distinct domains. We use $D_i$ to denote the domain membership for the $i$-th observation so that $D_i \in \{0, \ldots, G\}$ with $D_i = 0$ indicating the target domain. The unified data generating process for all domains can be expressed as

$$
Y_i|D_i = g \sim \text{Cat}(\pi^{(g)}),
$$

$$
Z_i|Y_i = c, D_i = g \sim \text{Cat}(\lambda_c^{(g)}),
$$

$$
X_{ij}|Y_i = c, Z_i = k \sim \text{Bern}(\theta_{ckj}).
$$

That is, we assume there are $K$ latent symptom profiles nested within each of the $C$ causes across all domains, but the distribution of the causes and latent classes are domain-specific. The domain-specific distributions of $p(X|Y)$ illustrated earlier can then be explained by different mixing over latent classes. Similar to the single-domain model, we put independent stick breaking priors on the domain-specific weights and Dirichlet priors for the CSMF, i.e., for $g = 1, \ldots, G$,

$$
\lambda_{ck}^{(g)} = V_{ck}^{(g)} \prod_{l<k} (1 - V_{cl}^{(g)}), \quad V_{ck}^{(g)} \sim \text{Beta}(1, \omega_{ck}^{(g)}),
$$

$$
\omega_c^{(g)} \sim \text{Gamma}(a_\omega, b_\omega), \quad \pi^{(g)} \sim \text{Dirichlet}(\alpha^{(g)}).
$$

When performing cause-of-death assignment for a target domain without labeled death, the estimation of the mixing weights $\lambda^{(0)}$ should ideally borrow information from the col-
lection of mixing weights \( \{ \lambda^{(g)} \}_{g=1,...,G} \) in the training domains. For the target domain, we let \( \lambda^{(0)}_{ck} = \sum_{g=1}^{G} \eta_{cg} \lambda_{ck}^{(g)} \), i.e., we assume the deaths due to each cause from the new domain can be modeled as a mixture of deaths due to the same cause from the observed domains. Compared to putting independent priors on \( \lambda^{(0)} \) and estimating it without constraints, this parameterization allows the cause-of-death classification to be less sensitive to small and noisy target domain data. We note that this modeling assumption can be relaxed if domain-level covariates are known and can be used to model domain similarities. Here we consider the more common case where no additional information about the target domain exists and leave covariate-based modeling to future work. Specifically, we consider the following two strategies to parameterize \( \lambda^{(0)} \), and a graphical model representation of the model is summarized in Figure 2.

1. **Domain-level mixture:** We fix \( \eta_c = \eta \) to be the same for each cause \( c = 1, ..., C \) and let \( \eta \sim \text{Dirichlet}(\alpha_\eta) \). That is, the relationship between the new mixing weights \( \lambda^{(0)} \) and the existing weights \( \{ \lambda^{(1)}, ..., \lambda^{(G)} \} \) do not vary across causes of death.

2. **Domain-cause-level mixture:** The simple domain-level mixture formulation can be too restrictive when the cause-of-death distributions are unbalanced, as the symptom distributions in the target domain are assumed to be similar to the domains with large \( \eta \) conditional on any cause, including those rarely observed in these domains. Instead, we can perform cause-specific partial pooling by letting \( \eta_c \sim \text{Dirichlet}(\alpha^{(c)}_\eta) \). We specify the concentration parameters to be \( \alpha^{(c)}_\eta = \alpha_\eta m_{cg} \), where \( m_{cg} \) is the fraction of deaths from domain \( g \) among all deaths due to cause \( c \). This domain-cause-level mixture encourages \( \lambda^{(0)}_{c} \) to be similar to the corresponding weights in domains with larger sample sizes for cause \( c \). When domain-level covariates exist, \( \alpha^{(c)}_\eta \) can also be parameterized as proportional to other domain-level similarity measures.
Figure 2: Graphical model of our Bayesian nested latent model. Boxes indicate replication of random variables and shaded nodes the observations. The middle block correspond to observed and latent variables for each death. The right block correspond to the shared parameters describing the symptom profile collections. The left block correspond to parameters describing the site-level mixing of the latent symptom profiles.

With a sufficient number of latent classes, both the single-domain and multi-domain models can capture the data distribution in the training data equally well. The advantage of the multi-domain model, however, is in the out-of-domain prediction, as $\lambda^{(0)}$ can be estimated with more flexibility. The similarity parameter, $\eta$, quantifies the similarity between the target and existing training domains in terms of the decomposition of latent classes within causes, as illustrated in Section 5.
3.3 Practical considerations

In the context of VA, $\theta$ serves as the role of symptom-cause-information (SCI) (Clark et al., 2018) on the latent space and is assumed to be shared across domains. In our implementation, we fix the distribution of $\theta$ to its posterior distribution given the training data $T$ when modeling the VAs collected in the target domain. This modular procedure has several advantages. First, it prevents the information on $\theta$ from flowing from the target domain to the training domains, as the target domain typically have few or no labeled deaths. This type of modularized procedure is known as ‘cutting the feedback’ (Plummer, 2015) and has been studied extensively in different applications (Lunn et al., 2009; Cucala et al., 2009; Zigler et al., 2013; Jacob et al., 2017; Pompe and Jacob, 2021, e.g.). In addition, a naive joint model likely will lead to the unlabeled deaths forming their own mixtures away from the latent classes in the training domains (Adams and Ghahramani, 2009). Second, as the information from the training domains can be summarized in the posterior distributions of $(\theta, \lambda)$, users of the algorithm do not need to have access to the entire training dataset, which could be subject to data sharing restrictions. Third, the modular approach allows us to improve the exploration of parameter space by running multiple chains in parallel in the training stage, leading to a better approximation of $p(\theta|T)$ and reduced computational cost for practitioners.

Lastly, data collected through VA inevitably contain a high proportion of missing responses due to the respondents not knowing whether certain symptoms exist or not being willing to talk to the interviewer about them. Some of the missing symptoms can be deterministically inferred by the logic of the questionnaire whereas most of them are difficult to impute systematically. We assume these missing responses are missing at random, i.e., the probability of the missing data mechanism depends on observed data but not on missing
values. This is a standard assumption made by existing VA algorithms (McCormick et al., 2016; Kunihama et al., 2020; Moran et al., 2021) and allows us to avoid building another imputation model, which is extremely challenging to specify and validate in practice. Under the missing at random assumption, we can conduct inference on model parameters using only the observed data. More discussions on this assumption can be found in Kunihama et al. (2020).

3.4 Related work outside of VA

Learning from data collected in different domains is an active area of research in machine learning and has been explored in various applications including natural language processing (Ramponi and Plank 2020), visual classification (Wang and Deng 2018), sentiment prediction (Glorot et al., 2011), and more recently in prediction problems in public health and clinical settings (Rehman et al., 2018; Mhasawade et al., 2020; Laparra et al., 2020). The task we consider in this paper is usually categorized as unsupervised domain adaptation (UDA), where no labeled data exist in the target domain. In particular, our estimation framework can be seen as a domain generalization method (Muandet et al., 2013) as the training stage does not involve the target domain data. A comprehensive review of UDA can be found in Wilson and Cook (2020). Most of the UDA methods rely on creating domain-invariant representations of the features (e.g., Tzeng et al., 2014; Shen et al., 2022). There is extensive literature focusing on using deep learning to extract such representation in computer vision and natural language processing, but these methods are less appropriate in the context of VA as they require extensive training data (Oquab et al., 2014). Instead, we take a flexible parametric approach and learn a parsimonious and interpretable symptom representation that is shared across domains.
Bayesian hierarchical models are routinely used to model heterogeneous, multi-domain data in many scientific fields, though less common in the domain adaptation literature. Bruzzone and Prieto (2001) and Raghuram et al. (2012) proposed Gaussian mixture models to learn the data representation of training domains and apply to target domains in remote sensing. Wood and Teh (2009) proposed nonparametric Bayesian methods for domain adaptation in language models. Hajiramezanali et al. (2018) and Boluki et al. (2021) developed Bayesian hierarchical models for modeling RNA sequencing data from multiple domains. While being tailored for different applications, the common conceptual framework is to model the data distribution in different domains as mixtures of shared latent clusters, an insight that we also use in building our proposed latent class model.

4 Posterior inference

We have the following posterior joint distribution given the training dataset $T = \{(x_i, y_i, d_i)\}_{i=1,\ldots,n}$, $p(\pi, \lambda, \gamma, \phi, \omega | T) \propto \prod_{i=1}^{n} p(y_i | \pi^{(d_i)}) \sum_{z_i=1}^{K} \left( p(z_i | y_i, \lambda^{(d_i)}) \prod_{j=1}^{p} p(x_{ij} | y_i, z_i, \gamma, \phi, \delta) \right)$

$$\times \prod_{c=1}^{C} \prod_{j=1}^{p} p(\gamma_{cj}) \prod_{c=1}^{C} \prod_{k=1}^{K} \prod_{j=1}^{p} p(\phi_{ckj}) p(\delta_{ckj} | \tau_c) \prod_{c=1}^{C} p(\tau_c)$$

$$\times \prod_{g=1}^{G} p(\pi^{(g)}) p(\lambda^{(g)} | \omega^{(g)}) p(\omega^{(g)}).$$

The posterior distribution of the single-domain model is a special case with $G = 1$ and thus omitted. In the target domain, the modularized joint posterior distribution is

$$p(Y^{(0)}, \pi^{(0)} | X^{(0)}, T) \propto \int \prod_{i=1}^{n} p(y_i^{(0)} | \pi^{(0)}) \sum_{z_i^{(0)}=1}^{K} \left( p(z_i^{(0)} | y_i^{(0)}, \lambda^{(0)}) \prod_{j=1}^{p} p(x_{ij}^{(0)} | y_i^{(0)}, z_i^{(0)}, \theta) \right)$$

$$\times p(\pi^{(0)}) p(\lambda^{(0)} | \lambda^{(1)}, \ldots, \lambda^{(G)}, \eta) p(\eta)$$

$$\times p(\lambda^{(1)}, \ldots, \lambda^{(G)}, \theta | T) d(\eta, \lambda^{(0)}, \ldots, \lambda^{(G)}, \theta).$$
The posterior distribution of the model parameters is not available in closed form, but we can easily draw posterior samples from a Gibbs sampler. The detailed procedure of fitting the model is described next.

### 4.1 Training stage

We first infer the parameters associated with the training data using the following steps.

1. Sample the latent class membership $Z_i$ for $i \in \mathcal{T}$ with
   \[
   p(Z_i = k|Y_i = c, D_i = g, \theta, \lambda) \propto \lambda_{ck} \prod_j \theta_{ckj}^{x_{ij}} (1 - \theta_{ckj})^{1-x_{ij}}.
   \]

2. Sample the stick-breaking parameters $V^{(g)}$ and $\omega^{(g)}$ for $g = 1, ..., G$. Denote $n_{ck}^{(g)} = \sum_i 1_{Y_i = c, Z_i = k, D_i = g}$. The full conditional is
   \[
   V_{ck}^{(g)}|Y, Z, \omega^{(g)} \sim \text{Beta}(1 + n_{ck}^{(g)}, \omega_{c}^{(g)} + \sum_{l>k} n_{cl}^{(g)}), \ c = 1, ..., C; \ k = 1, ..., K - 1,
   \]
   \[
   \omega_{c}^{(g)}|V^{(g)} \sim \text{Gamma}(a_\omega + K - 1, b_\omega - \sum_{k=1}^{K-1} \log(1 - V_{ck}^{(g)})), \ c = 1, ..., C.
   \]

3. Sample the response probabilities $\theta$ after integrating out the latent binary indicators $\delta$. The posterior conditional of $\theta_{ckj}$ is a mixture of Beta distribution and point mass at the baseline vector $\gamma_{cj}$. We let $n_{ckj1} = \sum_i 1_{Y_i = c, Z_i = k, X_{ij} = 1}$ and $n_{ckj0} = \sum_i 1_{Y_i = c, Z_i = k, X_{ij} = 0}$. The posterior conditional of $\theta_{ckj}$ is
   \[
   p(\theta_{ckj}|\cdot) = \tilde{\pi}_{ckj} \text{Beta}(1 + n_{ckj1}, \nu_\phi + n_{ckj0}) + (1 - \tilde{\pi}_{ckj}) I(\gamma_{cj}),
   \]
   where $I(\gamma_{cj})$ is the point mass at $\gamma_{cj}$. The mixing probability is
   \[
   \tilde{\pi}_{ckj} = \frac{\tau_c B(1 + n_{ckj1}, \nu_\phi + n_{ckj0})}{\tau_c B(1 + n_{ckj1}, \nu_\phi + n_{ckj0}) + (1 - \tau_c) B(1, \nu_\phi) \gamma_{cj}^{n_{ckj1}} (1 - \gamma_{cj})^{n_{ckj0}}},
   \]

4. Sample the latent binary indicators $\delta$ from its full conditional,
   \[
   \delta_{ckj} \sim \text{Bern}(\tilde{\pi}_{ckj}).
   \]
5. Sample the sparsity level parameter $\tau_c$,

$$
\tau_c \sim \text{Beta}(1 + \sum_j \sum_k \delta_{ckj}, \nu + \sum_j \sum_k (1 - \delta_{ckj})).
$$

6. Sample the baseline vector $\gamma$ from its full conditionals. Let $O_j = \{i, x_{ij} \text{ not missing}\}_{i=1,\ldots,n}$, we have

$$
\gamma_{cj}|X, Y, Z, \delta \sim \text{Beta}(a_\gamma + \sum_{i \in O_j} x_{ij}(1 - \delta_{yi,zij}), b_\gamma + \sum_{i \in O_j} (1 - x_{ij})(1 - \delta_{yi,zij})).
$$

7. Update the CSMF vector in each domain $\pi^{(g)}|Y$. Let $n_c^{(g)} = \sum_i 1_{Y_i = c, D_i = g}$,

$$
\pi^{(g)}|Y \sim \text{Dirichlet}(\alpha_{\pi,1} + n_1^{(g)}, \ldots, \alpha_{\pi,C} + n_C^{(g)}).
$$

When there exist observations from the training domains with unknown labels, the above posterior sampling procedure can be easily modified by updating the cause-of-death assignment and latent class indicator for each unlabeled data with

$$
p(Y_i = c, Z_i = k|X_i, D_i = 0, \lambda, \theta) \propto \pi_c^{(0)} \eta_{cg} \lambda_{ck} \prod_j (\theta_{ckj})^{x_{ij}} (1 - \theta_{ckj})^{1-x_{ij}}.
$$

### 4.2 Prediction stage

When performing cause-of-death assignment for the target dataset, we take the modular approach described in Section 3.3 by plugging in the posterior draws of $\theta$ and $\lambda$ from the training stage model. We describe the sampler for the other parameters in the multi-domain model below and leave the case of single-domain model to the Supplementary Materials.

1. Sample the cause-of-death assignments $Y_i$ and latent class indicator $Z_i$. To facilitate easier computation, we augment the data with latent indicator $\tilde{D}_i \in \{1, \ldots, G\}$ for observations from the target population and sample all latent indicators jointly with

$$
p(Y_i = c, Z_i = k, \tilde{D}_i = g|X_i, D_i = 0, \lambda, \theta) \propto \pi_c^{(0)} \eta_{cg} \lambda_{ck} \prod_j (\theta_{ckj})^{x_{ij}} (1 - \theta_{ckj})^{1-x_{ij}}.
$$
2. For the multi-domain model with domain-level mixture, update the domain similarity weights $\theta$ with

$$\theta | Y, \bar{D} \sim \text{Dirichlet}(\alpha_{\theta,1} + \sum_i 1_{\bar{D}_i=1}, \ldots, \alpha_{\theta,G} + \sum_i 1_{\bar{D}_i=G}).$$

For the multi-domain model with domain-cause-level mixture,

$$\theta_c | Y, \bar{D} \sim \text{Dirichlet}(\alpha_{\theta,1}m_{c1} + \sum_i 1_{Y_i=c,\bar{D}_i=1}, \ldots, \alpha_{\theta,Gm_{cG}} + \sum_i 1_{Y_i=c,\bar{D}_i=G}).$$

3. Update the CSMF vector in the target domain $\pi^{(0)}|Y$ with

$$\pi^{(0)}|Y \sim \text{Dirichlet}(\alpha_{\pi,1}^{(0)} + n_{1}^{(0)}, \ldots, \alpha_{\pi,C}^{(0)} + n_{C}^{(0)})$$

where $n_{c}^{(0)} = \sum_i 1_{Y_i=c,\bar{D}_i=0}$.

4.3 Computational consideration

The latent class model formulation leads to a multimodal posterior distribution that could be difficult to explore using MCMC in practice. One of the reasons is likelihood invariance to permutation of class labels. When we model the mixing weights in the target domain as a weighted average of the weights in the training domains, however, the model parameters for the target site do not depend on the labeling of latent classes, making the computation in the prediction stage invariant to the latent class labels. This allows us to implement the training stage sampler multiple times in parallel with different starting values and combine the posterior draws of $(\theta, \lambda)$ afterwards. To further improve the predictive performance, we adopt the stacking approach proposed by Yao et al. (2022) to combine the multiple MCMC chains.

Finally, We note that the sampler proposed in this section enjoys significant advantage over models that use latent Gaussian representations to characterize symptom dependence.
The classification stage of LCVA essentially reduces to allocating observations to $CK$ or $CKG$ latent classes with conditional independent response probabilities, which has the same order of computational complexity as the InSilicoVA algorithm (McCormick et al., 2016).

5 Results

In this section, we present results of our model evaluated on the PHMRC gold-standard VA dataset (Murray et al., 2011a). The PHMRC dataset has been used extensively in validating and comparing VA cause-of-death assignment methods (McCormick et al., 2016; Kunihama et al., 2020; Moran et al., 2021; Clark et al., 2018). It consists of 7,841 adult deaths collected from six study sites (Andhra Pradesh, India; Bohol, Philippines; Dar es Salaam, Tanzania; Mexico City, Mexico; Pemba Island, Tanzania; and Uttar Pradesh, India). All deaths occurred in health facilities and gold-standard causes are determined based on laboratory, pathology and medical imaging findings. The cause of death is coded into 34 categories, and we pre-process the raw dataset into 168 binary symptoms as described in McCormick et al. (2016). The data processing steps are implemented using the openVA package (Li et al., 2023).

We present three sets of experiments in this section. We first evaluate the out-of-domain predictive performance of different models for a leave-one-site-out experiment. To further explore how these methods behave under more extreme data shift, we then construct re-sampled datasets based on each target domain, and evaluate the models on these replicated datasets. Finally, we consider the case where some local labeled data exist in the target domain and compare different approaches to incorporate the labeled data. In addition to the evaluation of predictive performance, we also compare computational time of different
methods and discuss the estimated latent symptom profiles. We also perform comparisons on synthetic data and results of these simulations are included in the Supplementary Materials.

5.1 Prediction for a new study site

For our first evaluation, we consider the task of out-of-domain prediction for each of the six study sites. We iteratively take one site as the target domain and the rest of the five sites as training domains. A summary of the true CSMFs for each site is included in the Supplementary Materials.

Throughout all the experiments in this paper, we use the same six sets of training data, each including five sites. Since our goal is to develop models that are robust to unseen target domains, we should choose $K$ based solely on the training data. To determine the value of $K$, we first fit the multi-domain model with a large $K$ and truncate the latent class rarely occupied. For each experiment, we first let $K = 15$ and monitor the number of latent classes occupied in the posterior samples of $Z_i$. Figure 3 plots the fraction of posterior samples where each latent class is occupied for each cause of death in each experiment. The latent classes after the largest 10 classes are occupied in less than 5% of posterior samples across almost all causes and experiments. Therefore, we let $K = 10$ in all our subsequent analysis of the PHMRC data. The performance of LCVA does not vary much from different choices of $K$. Additional sensitivity analysis are provided in the Supplementary Materials.

We consider both the multi-domain (LCVA-M) models in Section 3.2 and the single-domain (LCVA-S) models in Section 3.1 by ignoring the domain indicators in the training data. For the concentration parameters, we let $\alpha_\pi = \alpha_\pi^{(0)} = \alpha_q = 1$. In the training
Figure 3: The fraction of latent classes utilized for each experiment. The latent classes are sorted by the fraction of posterior draws where they are occupied. Each line represents one cause of death and each panel represents a training dataset where the target site in the label is excluded.

stage, we ran six parallel MCMC chains with different starting values for 4,000 iterations and discarded the first 1,000 posterior draws as burn-in. For LCVA-S with new weights, the posterior draws from a single chain were used in the prediction stage. For the other models, we re-weighted the six chains using stacking (Yao et al., 2022) and generated 4,000 posterior samples from the combined posterior distribution. Finally, the first half of the chain in the prediction stage were discarded as burn-in.

We evaluated the performance of the proposed model based on the classification accuracy of the predicted top cause of death and the so-called ‘CSMF accuracy’ (Murray et al., 2011b), a widely used metric to compare the estimated CSMF vector with the truth. The CSMF accuracy metric is a form of the normalized absolute error (González et al., 2017).
It measures the $L_1$ distance between the estimated and true CSMF, and is normalized to range between 0 (worst) and 1 (best). It is defined as \( \text{CSMF}_{\text{acc}}(\hat{\pi}) = 1 - \frac{\sum_{c=1}^{C} |\hat{\pi}_c - \pi_c|}{2(1 - \min_c \pi_c)} \), where \( \pi_c \) is the true CSMF calculated by the empirical distribution in the test data. Unlike the top cause accuracy that focuses on individual level performance, CSMF accuracy evaluates the task of prevalence estimation, which is more informative for VA practitioners, especially in unbalanced datasets.

We benchmarked the performance of the proposed model to the InSilicoVA algorithm. InSilicoVA is one of the most widely adopted VA methods used in practice and has been shown to have better or comparable performance compared to other VA models in use (McCormick et al., 2016; Clark et al., 2018). We also compared with the Bayesian factor model (Kunihama et al., 2020) and the FARVA model (Moran et al., 2021). Both models approximate the joint distribution of symptoms using a latent Gaussian model and have been shown to significantly improve the predictive performance compared to InSilicoVA.

For the FARVA model, we included a binary covariate indicating age at death exceeding 65. The inclusion of covariates did not change the performance metric of FARVA by much in our experiments. We follow the recommendation by the authors of these methods in setting the tuning parameters and length of MCMC iterations. For the InSilicoVA algorithm, we ran the MCMC for 4,000 iterations. For the Bayesian factor model, we used 6 latent factors and obtained 1,000 posterior samples after discarding the first 500 samples as burn-in and saving every fifth sample. Far FARVA, we used 10 latent factors and obtained 500 posterior samples after discarding the first 2,000 samples as burn-in and saving every tenth sample.

Figure 4 summarizes the top cause accuracy and CSMF accuracy for the different models evaluated on the six sites respectively. All models show significantly improved top cause accuracy compared to InSilicoVA. The CSMF accuracy of the three classes of models are...
also comparable and improves from the InSilicoVA results except in two cases (FARVA in Bohol and LCVA in Andhra Pradesh). There is no model that dominates either predictive metrics in all six experiments.

We note, however, that an important caveat in this experiment is that the CSMFs in most sites are similar in the PHMRC dataset, except for Pemba. A comparison of the empirical CSMFs are included in the Supplementary Materials. Both the Bayesian factor model and FARVA assume a common data generating process for causes of deaths in both training and target datasets with a shared parameter $\pi$, and thus enjoy stronger regularization from the prior compared to LCVA. However, such model formulation can negatively affect the predictive performance in domains with a very different cause-of-death distribution compared to the training domains, such as Pemba in this experiment. We further evaluate the performance of the cause-of-death assignment models under more extreme scenarios in the next subsection.

5.2 Prediction for a new study site under more extreme data shift

In order to evaluate the models under more extreme data shift, for each leave-one-site-out experiment, we created 50 synthetic target dataset by first generating $\pi^{(0)} \sim \text{Dirichlet}(0.1, ..., 0.1)$ and resampled deaths in the held-out site with replacement to match the generated target CSMF. This data generating process leads to the target CSMFs being concentrated on a small number of causes and thus very different from the overall prevalence in the training domains. This reflects more closely the cause-of-death distributions in small- and moderate-scale VA studies in practice. We again compare our methods to InSilicoVA, Bayesian factor model, and FARVA. The top cause accuracy and CSMF accuracy metrics vary significantly
Figure 4: Individual-level top cause prediction accuracy (left) and CSMF accuracy (right) of different models. The cells are colored by the relative difference from the InSilicoVA results, where blue cells indicate higher accuracy and red cells indicate lower accuracy. All variations of LCVA demonstrate good performance that is significantly higher than InSilicoVA except in one experiment. Bayesian factor model and FARVA also show comparable improvements.

across replications due to the high variability of the target domain CSMF. Thus we focus on the relative performances of the models. Figure 5 and 6 shows the pairwise comparison of all models in terms of the fraction of times one model outperforms the other in the two metrics across the six collection of target datasets. The simplest model, InSilicoVA, led to worse top cause accuracy estimation in almost all pairwise comparisons. However, in terms of the CSMF accuracy, which is more of interest for domains with unbalanced cause-of-death distribution, both Bayesian factor model and FARVA performed worse than InSilicoVA on more than half of the resampled datasets in four out of six sites. LCVA-M with domain-cause-level mixture is the only method that outperformed InSilicoVA in both metrics on more than half of the resampled datasets across all six sites, and it outperforms all three existing methods in CSMF accuracy on more than half of the resampled datasets. LCVA-M with domain-level mixture also perform very similarly. This demonstrates the
robustness of the proposed models to extreme data shift.

Figure 5: Pairwise comparison of all models in terms of top cause accuracy. The color of each cell represents the proportion of target datasets where the model on the left achieves higher top cause accuracy than the model on the bottom.

We further take a closer examination of the scale of the relative improvements of the proposed models. Figure 7 shows the percentage improvement of each model $m$ compared to the Bayesian factor model, i.e., $(\text{Acc}_m - \text{Acc}_{BF})/\text{Acc}_{BF}$ for both metrics. In Dar es Salaam and Pemba island, LCVA-M had similar performances compared to the Bayesian factor model. In the other four sites, LCVA-M models achieved around 10% to 40% in median improvements for top cause accuracy and 10% to 20% for CSMF accuracy. FARVA also outperformed the Bayesian factor model in five sites in terms of individual-level prediction, but showed worse CSMF accuracy across all sites. Additional comparisons are included in
Figure 6: Pairwise comparison of all models in terms of CSMF accuracy. The color of each cell represents the proportion of target datasets where the model on the left achieves higher CSMF accuracy than the model on the bottom.

Based on these two set of experiments and the synthetic data evaluations in the Supplementary Materials, LCVA-M with domain-level mixture was shown to be more robust to different data shift scenarios while enjoying the more parsimonious parameterization than the domain-cause-level mixture model. Thus we focus our attention on this model in the rest of the paper.
Figure 7: Boxplot of improvement over the Bayesian factor model (Kunihama et al., 2020) in top cause accuracy (left) and CSMF accuracy (right) by different models on 50 resampled target domain datasets based on deaths from each of the six sites.

5.3 Computation time

The proposed model is highly efficient in computation compared to the Bayesian factor model and FARVA. Across the six leave-one-site-out prediction experiments, the InSilicoVA algorithm is the fastest to implement as the symptoms are treated as conditionally independent. It required 20 seconds to run 1,000 iterations of MCMC. For the LCVA with domain-level mixture, every 1,000 iterations of MCMC required 142 seconds in the training stage, and 42 seconds in the prediction stage. In comparison, for the same tasks and 1000 iterations of the MCMC, the Bayesian factor model used 4,275 seconds and the FARVA model used 17,422 seconds. Data analysis using these two methods takes hours or days to finish, which is clearly infeasible as routine tools in resource-constrained settings. The
LCVA requires the least additional computational cost among the methods that take into account symptom dependence. The computation time are all evaluated on a MacBook Pro with 2.6 GHz 6-Core Intel Core i7 processor and 32 GB memory.

5.4 Interpreting the latent symptom profiles

We now turn to the estimated latent classes and domain similarity parameters. As an example, we examine the experiment treating Pemba island as the target domain and the other five larger sites as training domains. As an illustration, we focus on the multi-domain LCVA with domain-level mixture model. It is worth noting that the estimated latent classes are subject to permutation and further research is needed to relate latent classes with observable characteristics of the causes of death. Nevertheless, they still reveal interesting latent structures in these data. The top panel of Figure 8 shows the estimated conditional probabilities, $\theta$. We sort the symptoms by their largest conditional probabilities given a latent class, i.e., $\max_k \theta_{ckj}$, and show the conditional probabilities of the top 20 symptoms for the six most prevalent causes in Pemba. Within each cause of death, most of the symptoms have similar conditional probabilities across latent classes, as encouraged by the prior on $\theta$. Taking deaths due to ‘other Non-Communicable Diseases (NCD)’ as an example, the latent symptom profiles differ most significantly on the response probabilities for the sex of the respondent, injury, and specific symptoms such as convulsion, chest pain, having trouble breathing, and cough. The corresponding mixing weights are shown in the lower left panel of Figure 8. Mexico city and Dar es Salaam have more different latent class decomposition than the other sites. For example, comparing deaths due to other NCD in Dar es Salaam with the other sites, the 1st and 3rd latent classes were utilized less often compared to the 2nd and 4th, indicating that symptoms such as having
trouble breathing are less likely to be present for deaths due to other NCDs compared to other sites. We note, however, that such differences could be due to either different decomposition of specific NCDs or differential reporting across sites, and more studies are needed to understand their underlying mechanism. We also observe that for simpler causes such as drowning and falling, the latent symptom profiles are heavily shrunk towards the baseline values, indicating that the conditional independence simplification being a useful approximation for these causes. The shrinkage protects the model from overfitting on the noise in the training domains when \( K \) is large for a relatively simpler causes. Finally, the estimated \( \eta \) in the lower right panel of Figure 8 shows how \( \lambda^{(0)} \) is related to the existing mixing weights through the similarity parameter \( \eta \). The domain with the most similar mixing weights to Pemba is estimated to be Dar es Salaam, which is as expected given both sites are in Tanzania.

5.5 Incorporating labeled data from the target domain

Finally, we consider the situation where a small subset of labeled data is available in the target domain. In this case, the target population \( T_0 = L \cup U \), where \( L \) are the data with labels and \( U \) are the data without labels. The LCVA model can incorporate such information naturally by modeling all deaths in \( T_0 \) and skipping the sampling of \( Y_i \) for the deaths in \( L \). This leads to a joint modeling approach in the prediction stage in a semi-supervised fashion. An alternative strategy is to calibrate the estimated CSMF to the known labels after fitting a model on only the training data and use \( L \) to estimate the misclassification matrix and de-bias the estimated \( p_{T_0}(Y) \) (Fiksel et al., 2021).

Here we again consider the six experiments of out-of-domain prediction in Section 5.1. For the calibration step, we took the posterior means of the estimated individual cause-of-
Figure 8: Top: Posterior means of the conditional probabilities of observing the selected subset of symptoms given the top six causes of death in Pemba and latent class membership. Bottom: Posterior means of the mixing weights of latent classes by domain (left) and the posterior distribution of the similarity parameter $\eta$. Pemba island is the target domain. The mixing weights corresponding to drowning in Mexico city are removed since there is no drowning deaths in Mexico city in this dataset.
death probabilities from InSilicoVA and LCVA-M with domain-level mixture as the input to the calibration procedure in [Fiksel et al. (2021)], and randomly selected 30% of deaths in the target domain to reveal the true labels. For each randomly selected set of labeled deaths, we compared the three estimates of the CSMFs: calibrated InSilicoVA, calibrated LCVA, and LCVA with partial labels treated as known in the prediction stage. The results over 50 randomly sampled $L$ are summarized in Figure 9. With the additional information of the known labels, the CSMF accuracy improves for both InSilicoVA and LCVA. By directly incorporating the deaths with known causes in the joint modeling approach, we were able to obtain the highest CSMF accuracy in all six sites using LCVA.

We note, however, that the calibration approach in [Fiksel et al. (2021)] were developed for a much smaller cause list. The calibration procedure is less effective in this analysis due to the need of estimating a $34 \times 34$ dimensional error transition matrix using a small number of labeled data. In applications where only broader cause categories are of interest, the calibration procedure can sometimes perform similarly well compared to the joint modeling approach. We experimented with calibration using aggregated 11 and 5 broader causes in the Supplementary Materials.

6 Discussion

Quantifying the population distribution of causes of death and assigning causes to specific deaths using VAs are challenging, especially when using information derived from non-local data sources. In this paper, we proposed a statistical framework for modeling VA data collected over multiple domains and quantifying the cause-of-death distributions for deaths in a new domain. The proposed model framework uses a parsimonious nested latent class model to characterize the joint distributions of symptoms given causes, which allows
Figure 9: Comparison of LCVA-M with domain-level mixture and InSilicoVA when the cause of death of 30% of data in the target domain are known. The horizontal line indicates the CSMF accuracy of the two methods before calibrating to the labeled deaths. The box plot shows the distribution of the CSMF accuracy over 50 resampled labeled dataset.

The proposed method could be extended in a few different ways. First, we have treated the domains as independent datasets in this paper. In many situations, data are collected over different sub-national regions and over time. Structural information about spatial and temporal correlation, as well as domain-level similarities could further improve the modeling of domain-specific distributions, as cause of death and latent class distributions...
are likely to vary smoothly across similar domains. Summarizing domain heterogeneity and similarities using the latent class model can also be of scientific interest to understand the underlying population. In addition, symptom distributions conditional on similar causes of death may also show similarities. Tree-based methods for domain adaptation can be a fruitful direction to explore such additional information (Wu et al., 2021). Second, it is an important question to evaluate the discriminative power of VA questions in assigning causes, and identify the questions that do no contribute to cause assignments across different domains. This can facilitate the design of shorter VA questionnaires in the future. The proposed framework can be extended to perform symptom selection by further shrinking the response probabilities across different causes of death to the same values. Third, while our model focuses on the prediction task, the discovery and inference of interpretable latent classes shared across different populations alone is a topic of significant practical interest (De Vito et al., 2019). It is a challenging task in the VA context, as the collected symptoms could be measured inconsistently across domains and subject to reporting biases. This is related to the discussion of transportability in the causal inference literature (Wu et al., 2019; Ackerman et al., 2019).

Finally, we conclude by highlighting two additional open questions. First, the proposed model relies on reference deaths with known causes of death from multiple training domains. A common practice of collecting labeled VA data is by having physician experts read and assign a cause best describing the death. The assigned causes of death are treated as the ground truth. However, many factors could contribute to the misclassification of deaths and lead to noise in these training labels. In particular, when a new disease emerges and limited information about the disease is known, such as during the early periods of the COVID-19 pandemic, higher rate of misclassification may be expected from the training labels. A
key challenge, therefore, will be to identify latent classes that correspond to misclassified or unseen causes. Methods dealing with label noises in classification tasks (Liu and Guo, 2020; Liu, 2021) could potentially improve cause-of-death assignment models and produce more realistic characterization of uncertainties. Second, a grand challenge with VA is to embed the analysis framework into the mortality surveillance system to detect changes and abnormalities. Modeling the temporal drift of the data distribution(3,3),(985,993) is highly useful for surveillance purposes. Covariate-dependent modeling (Moran et al., 2021) is also important in this context, as additional information from both the population mortality surveillance and individual deaths could further help understand transferability and explain changes in mortality profile. We leave these topics for future work.

References

Ackerman, B., Siddique, J., and Stuart, E. A. (2019). Transportability of outcome measurement error correction: from validation studies to intervention trials. *arXiv preprint arXiv:1907.10722*.

Adams, R. P. and Ghahramani, Z. (2009). Archipelago: nonparametric Bayesian semi-supervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1–8.

Blanco, A., Perez, A., Casillas, A., and Cobos, D. (2020). Extracting cause of death from verbal autopsy with deep learning interpretable methods. *IEEE Journal of Biomedical and Health Informatics*.

Boluki, S., Qian, X., and Dougherty, E. R. (2021). Optimal Bayesian supervised domain adaptation for RNA sequencing data. *Bioinformatics*, 37(19):3212–3219.
Breiman, R. F., Blau, D. M., Mutevedzi, P., Akelo, V., Mandomando, I., Ogbuanu, I. U., Sow, S. O., Madrid, L., El Arifeen, S., Garel, M., et al. (2021). Postmortem investigations and identification of multiple causes of child deaths: An analysis of findings from the Child Health and Mortality Prevention Surveillance (CHAMPS) network. *PLoS medicine, 18*(9):e1003814.

Bruzzone, L. and Prieto, D. F. (2001). Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing, 39*(2):456–460.

Byass, P., Chandramohan, D., Clark, S. J., D’ambruoso, L., Fottrell, E., Graham, W. J., Herbst, A. J., Hodgson, A., Hounton, S., Kahn, K., et al. (2012). Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool. *Global health action, 5*(1):19281.

Byass, P., Huong, D. L., and Van Minh, H. (2003). A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in Vietnam. *Scandinavian Journal of Public Health, 31*(62 suppl):32–37.

Byass, P., Hussain-Alkhateeb, L., D’Ambruoso, L., Clark, S., Davies, J., Fottrell, E., Bird, J., Kabudula, C., Tollman, S., Kahn, K., et al. (2019). An integrated approach to processing who-2016 verbal autopsy data: the interva-5 model. *BMC Medicine, 17*(1):1–12.

Clark, S. J., Li, Z. R., and McCormick, T. H. (2018). Quantifying the contributions of training data and algorithm logic to the performance of automated cause-assignment algorithms for Verbal Autopsy. *arXiv: 1803.07141.*
Cucala, L., Marin, J., Robert, C., and Titterington, D. (2009). Bayesian inference in k-nearest-neighbour classification models. *J. American Statist. Assoc*, 104(485):263–273.

Datta, A., Fiksel, J., Amouzou, A., and Zeger, S. L. (2021). Regularized Bayesian transfer learning for population level etiological distributions. *Biostatistics*, 22(4):836–857.

Daume III, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2019). Multi-study factor analysis. *Biometrics*, 75(1):337–346.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Fiksel, J., Datta, A., Amouzou, A., and Zeger, S. (2021). Generalized Bayes quantification learning under dataset shift.

Flaxman, A. D., Vahdatpour, A., Green, S., James, S. L., and Murray, C. J. (2011). Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Population health metrics*, 9(1):29.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513—-520.

González, P., Castaño, A., Chawla, N. V., and Coz, J. J. D. (2017). A review on quantification learning. *ACM Computing Surveys (CSUR)*, 50(5):1–40.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.
Hajiramezanali, E., Zamani Dadaneh, S., Karbalayghareh, A., Zhou, M., and Qian, X. (2018). Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. *Advances in Neural Information Processing Systems, 31.*

Jacob, P. E., Murray, L. M., Holmes, C. C., and Robert, C. P. (2017). Better together? statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719.*

King, G. and Lu, Y. (2008). Verbal autopsy methods with multiple causes of death. *Statistical Science, 100*(469).

Kunihama, T., Li, Z. R., Clark, S. J., and McCormick, T. H. (2020). *The Annals of Applied Statistics, 14*(1):241–256.

Laparra, E., Bethard, S., and Miller, T. A. (2020). Rethinking domain adaptation for machine learning over clinical language. *JAMIA open, 3*(2):146–150.

Li, Z. R., McCormick, T. H., and Clark, S. J. (2020). Using Bayesian latent Gaussian graphical models to infer symptom associations in verbal autopsies. *Bayesian Analysis, 15*(3):781–807.

Li, Z. R., Thomas, J., Choi, E., McCormick, T. H., and Clark, S. J. (2023). The openVA toolkit for verbal autopsies. *The R Journal, pages 316–334.*

Liu, Y. (2021). Understanding instance-level label noise: Disparate impacts and treatments. In *International Conference on Machine Learning, pages 6725–6735. PMLR.*

Liu, Y. and Guo, H. (2020). Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning, pages 6226–6236. PMLR.*
Lozano, R., Lopez, A. D., Atkinson, C., Naghavi, M., Flaxman, A. D., and Murray, C. J. (2011). Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics, 9*(1):1–13.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009). Combining mcmc with ‘sequential’ PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics, 36*(1):19–38.

Maher, D., Biraro, S., Hosegood, V., Isingo, R., Lutalo, T., Mushati, P., Ngwira, B., Nyirenda, M., Todd, J., and Zaba, B. (2010). Translating global health research aims into action: the example of the alpha network. *Tropical Medicine & International Health, 15*(3):321–328.

McCormick, T. H., Li, Z. R., Calvert, C., Crampin, A. C., Kahn, K., and Clark, S. J. (2016). Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association, 111*(515):1036–1049.

Mhasawade, V., Rehman, N. A., and Chunara, R. (2020). Population-aware hierarchical Bayesian domain adaptation via multi-component invariant learning. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 182–192.

Miasnikof, P., Giannakeas, V., Gomes, M., Aleksandrowicz, L., Shestopaloff, A. Y., Alam, D., Tollman, S., Samarikhalaj, A., and Jha, P. (2015). Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC Medicine, 13*(1):1.

Moran, K. R., Turner, E. L., Dunson, D., and Herring, A. H. (2021). Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data. *Journal of the Royal Statistical Society: Series C (Applied Statistics).*
Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530.

Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR.

Murray, C. J., Lopez, A. D., Black, R., Ahuja, R., Ali, S. M., Baqui, A., Dandona, L., Dantzer, E., Das, V., Dhingra, U., et al. (2011a). Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Population health metrics*, 9(1):27.

Murray, C. J., Lozano, R., Flaxman, A. D., Vahdatpour, A., and Lopez, A. D. (2011b). Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Population Health Metrics*, 9(1):28.

Nkengasong, J., Gudo, E., Macicame, I., Maunze, X., Amouzou, A., Banke, K., Dowell, S., and Jani, I. (2020). Improving birth and death data for african decision making. *The Lancet Global Health*, 8(1):e35–e36.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724.

Plummer, M. (2015). Cuts in Bayesian graphical models. *Statistics and Computing*, 25(1):37–43.
Pompe, E. and Jacob, P. E. (2021). Asymptotics of cut distributions and robust modular inference using posterior bootstrap. *arXiv preprint arXiv:2110.11149*.

Raghuram, J., Miller, D. J., and Kesidis, G. (2012). Semisupervised domain adaptation for mixture model based classifiers. In *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

Ramponi, A. and Plank, B. (2020). Neural unsupervised domain adaptation in NLP – a survey. *arXiv preprint arXiv:2006.00632*.

Rehman, N. A., Aliapoulios, M. M., Umawani, D., and Chunara, R. (2018). Domain adaptation for infection prediction from symptoms based on data from different study designs and contexts. *arXiv preprint arXiv:1806.08835*.

Sankoh, O. and Byass, P. (2012). The INDEPTH network: filling vital gaps in global epidemiology. *International Journal of Epidemiology*, 41(3).

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On causal and anticausal learning. In *International Conference on Machine Learning*, page 459–466. PMLR.

Serina, P., Riley, I., Stewart, A., James, S. L., Flaxman, A. D., Lozano, R., Hernandez, B., Mooney, M. D., Luning, R., Black, R., et al. (2015). Improving performance of the Tariff method for assigning causes of death to verbal autopsies. *BMC Medicine*, 13(1):1.

Shen, K., Jones, R., Kumar, A., Xie, S. M., HaoChen, J. Z., Ma, T., and Liang, P. (2022). Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2204.00570*. 

43
Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

Storkey, A. (2009). When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Wang, M. and Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Wilson, G. and Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46.

Wood, F. and Teh, Y. W. (2009). A hierarchical nonparametric bayesian approach to statistical language model domain adaptation. In *Artificial Intelligence and Statistics*, pages 607–614. PMLR.

World Health Organization (2021). WHO civil registration and vital statistics strategic implementation plan 2021-2025.

Wu, X., Braun, D., Kioumourtzoglou, M.-A., Choirat, C., Di, Q., and Dominici, F. (2019). Causal inference in the context of an error prone exposure: air pollution and mortality. *The annals of applied statistics*, 13(1):520.

Wu, Z., Li, Z. R., Chen, I., and Li, M. (2021). Tree-informed Bayesian multi-source domain adaptation: cross-population probabilistic cause-of-death assignment using verbal autopsy. *arXiv preprint arXiv:2112.10978*.
Yao, Y., Vehtari, A., and Gelman, A. (2022). Stacking for non-mixing bayesian computations: The curse and blessing of multimodal posteriors. *Journal of Machine Learning Research*, 23(79):1–45.

Zhou, J., Bhattacharya, A., Herring, A. H., and Dunson, D. B. (2015). Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*, 110(512):1562–1576.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). Model feedback in bayesian propensity score estimation. *Biometrics*, 69(1):263–273.
A Sampler for the single-domain model

1. Sample the cause-of-death assignments $Y_i$ and latent class indicator $Z_i$. To facilitate easier computation, we augment the data with latent indicator $\tilde{D}_i \in \{1, \ldots, G\}$ for observations from the target population and sample all latent indicators jointly with

$$p(Y_i = c, Z_i = k | X_i, D_i = 0, \lambda, \theta) \propto \pi_c^{(0)} \lambda_{ck} \prod_j (\theta_{ckj})^{x_{ij}} (1 - \theta_{ckj})^{1-x_{ij}}.$$ 

2. For the case with new mixing weights, we need to sample the target-specific $\lambda^{(0)}$, which is the same as Step 2 of the training stage model for $g = 0$.

3. Update the CSMF vector in the target domain $\pi^{(0)} | Y$ with

$$\pi^{(0)} | Y \sim \text{Dirichlet}(\alpha_{\pi,1}^{(0)} + n_{1}^{(0)}, \ldots, \alpha_{\pi,C}^{(0)} + n_{C}^{(0)})$$

where $n^{(0)}_c = \sum_i 1_{Y_i = c, D_i = 0}$.

B Synthetic data simulation

We simulate deaths from 5 training domains with known causes-of-death and one target domain without labeled data. We let $C = 20$ and $S = 50$ and 2,000 deaths in each domain.

We consider two distributions of causes in the training domains: (1) balanced cause-of-death distribution with $\pi^{(g)} \sim \text{Dirichlet}(5)$ and (2) unbalanced cause-of-death distribution with $\pi^{(g)} \sim \text{Dirichlet}(0.5)$. In each case, we consider four different data generating processes for $\lambda^{(g)}$:

1. **Single domain with conditional independent symptoms**: We let $K = 1$ and $\lambda^{(g)}_{c1} = 1$. That is, we assume the distribution of $p(X, Y)$ is the same in all the five domains and that the symptoms are conditionally independent given the causes;
2. **Single domain:** We let $\lambda^{(g)} = \lambda$ for all sites with $K = 10$ latent classes and generate $\lambda_c \sim \text{Dirichlet}(1)$ for each cause of death independently;

3. **Independent domains:** We let $\lambda_c^{(g)} \sim_{iid} \text{Dirichlet}(0.1)$ for all training and target domains. The Dirichlet prior encourages the weights to concentrate on a small number of latent classes. The target domain could contain latent classes that are rarely represented in the training domains in this case;

4. **Related domains:** We consider the case that two of the training domains are related to the target domain. More specifically, let $\lambda_c^{(g)} \sim_{iid} \text{Dirichlet}(0.1)$ for all training domains. For the target domain we generate $\lambda_c^{(0)}$ as the average of two randomly picked training sites with $\lambda_c^{(0)} = 0.5\lambda_c^{(g_1)} + 0.5\lambda_c^{(g_2)}$ for all $c$.

5. **Related domains with cause-specific association:** We consider the case where the training domains are related to the target domain and such association vary across causes of death. More specifically, let $\lambda_c^{(g)} \sim_{iid} \text{Dirichlet}(0.1)$ for all training domains, and $\lambda_c^{(0)} = \sum_g \eta_{cg} \lambda_c^{(g)}$ where $\eta_c \sim \text{Dirichlet}(0.5)$.

Since in the real data we expect the distribution of symptoms to be bimodal with heavier concentration around 0, we generated the response probabilities independently from a mixture distribution $\theta_{ckj} \sim 0.9\text{Beta}(1, 5) + 0.1\text{Beta}(5, 1)$. We further randomly mask 30% of the symptoms as missing.

In each scenario, we fit the four versions of the proposed latent class model with $K = 10$ using the same setup of MCMC iterations as described in the main paper for the PHMRC data experiments. To benchmark the performance with VA algorithms used in practice, we compare these metrics with InSilicoVA, which assumes symptoms are conditionally independent.
Figure [10] shows the performance of the algorithms. For case (1), all models achieve almost perfect prediction. The performance of InSilicoVA drops significantly in the latter four cases where the symptoms are dependent. Overall, the performance of the four variations of LCVA are consistently better than InSilicoVA. Within the four parameterizations of LCVA, the single-domain model with constant weights achieves the best results when data are truly generated from the same domain without heterogeneity. Multi-domain model with domain-level mixture achieves the best performance in case (4) where the target domain consists of mixtures of latent classes from two existing domains, with overall high prediction accuracy in both metrics across all experiments.

C Additional results on the PHMRC dataset

In this section, we present additional details and results for the analysis of the PHMRC dataset discussed in the main manuscript.

C.1 Site-specific CSMFs in the PHMRC dataset

Figure [11] shows the empirical distribution of the causes in the six study sites in the PHMRC data, which we treat as the true CSMFs in our evaluation.

C.2 Sensitivity analysis of $K$ on the leave-one-site-out experiment

Figure [12] to [15] summarize the top cause accuracy and CSMF accuracy for the different models evaluated on the six sites when $K = 3, 4, 5, \text{ and } 15$ respectively. The results are similar to the figure presented in the main paper with $K = 10$. 
Figure 10: CSMF accuracy (top row) and individual cause-of-death classification accuracy (bottom row) for the four simulation scenarios. LCVA performs significantly better than InSilicoVA. LCVA-S with constant weights achieves the best performance when data come from a single domain and LCVA-M with domain-level mixture achieves the best performance in the three cases with multiple domains.
Figure 11: True CSMFs in the six study sites in the PHMRC data.
Figure 12: Individual-level top cause prediction accuracy (left) and CSMF accuracy (right) of different models. The proposed models are fitted with $K = 3$. The cells are colored by the relative difference from the InSilicoVA results, where blue cells indicate higher accuracy and red cells indicate lower accuracy. All variations of LCVA demonstrate good performance that is significantly higher than InSilicoVA except in one experiment. Bayesian factor model and FARVA also show comparable improvements.

Figure 13: Individual-level top cause prediction accuracy (left) and CSMF accuracy (right) of different models. The proposed models are fitted with $K = 4$. The cells are colored by the relative difference from the InSilicoVA results, where blue cells indicate higher accuracy and red cells indicate lower accuracy. All variations of LCVA demonstrate good performance that is significantly higher than InSilicoVA except in one experiment. Bayesian factor model and FARVA also show comparable improvements.
|                           | Mexico | AP   | UP   | Dar  | Bohol | Pemba |
|---------------------------|--------|------|------|------|-------|-------|
| IndSilicoVA               | 0.23   | 0.33 | 0.24 | 0.27 | 0.27  | 0.28  |
| Bayesian Factor Model     | 0.23   | 0.37 | 0.39 | 0.33 | 0.32  | 0.4   |
| FARVA                     | 0.32   | 0.4  | 0.34 | 0.34 | 0.32  | 0.4   |
| LCVA-S: constant weights  | 0.28   | 0.34 | 0.38 | 0.34 | 0.34  | 0.45  |
| LCVA-S: new weights       | 0.27   | 0.3  | 0.34 | 0.33 | 0.32  | 0.4   |
| LCVA-M: domain-level mixture | 0.27 | 0.38 | 0.39 | 0.33 | 0.33  | 0.47  |
| LCVA-M: domain-cause-level mixture | 0.28 | 0.37 | 0.39 | 0.32 | 0.33  | 0.44  |

|                           | Mexico | AP   | UP   | Dar  | Bohol | Pemba |
|---------------------------|--------|------|------|------|-------|-------|
| IndSilicoVA               | 0.64   | 0.73 | 0.55 | 0.65 | 0.67  | 0.42  |
| Bayesian Factor Model     | 0.79   | 0.82 | 0.66 | 0.75 | 0.78  | 0.57  |
| FARVA                     | 0.77   | 0.79 | 0.66 | 0.67 | 0.64  | 0.55  |
| LCVA-S: constant weights  | 0.75   | 0.68 | 0.68 | 0.7  | 0.75  | 0.61  |
| LCVA-S: new weights       | 0.75   | 0.65 | 0.66 | 0.65 | 0.71  | 0.56  |
| LCVA-M: domain-level mixture | 0.78  | 0.74 | 0.73 | 0.67 | 0.78  | 0.65  |
| LCVA-M: domain-cause-level mixture | 0.77 | 0.72 | 0.7 | 0.67 | 0.75  | 0.63  |

Figure 14: Individual-level top cause prediction accuracy (left) and CSMF accuracy (right) of different models. The proposed models are fitted with $K = 5$. The cells are colored by the relative difference from the InSilicoVA results, where blue cells indicate higher accuracy and red cells indicate lower accuracy. All variations of LCVA demonstrate good performance that is significantly higher than InSilicoVA except in one experiment. Bayesian factor model and FARVA also show comparable improvements.

|                           | Mexico | AP   | UP   | Dar  | Bohol | Pemba |
|---------------------------|--------|------|------|------|-------|-------|
| IndSilicoVA               | 0.23   | 0.33 | 0.24 | 0.27 | 0.27  | 0.28  |
| Bayesian Factor Model     | 0.23   | 0.37 | 0.39 | 0.33 | 0.32  | 0.4   |
| FARVA                     | 0.32   | 0.4  | 0.34 | 0.34 | 0.32  | 0.4   |
| LCVA-S: constant weights  | 0.27   | 0.34 | 0.38 | 0.34 | 0.34  | 0.44  |
| LCVA-S: new weights       | 0.25   | 0.33 | 0.37 | 0.35 | 0.33  | 0.42  |
| LCVA-M: domain-level mixture | 0.28  | 0.36 | 0.4  | 0.33 | 0.34  | 0.46  |
| LCVA-M: domain-cause-level mixture | 0.28 | 0.35 | 0.39 | 0.32 | 0.34  | 0.45  |

|                           | Mexico | AP   | UP   | Dar  | Bohol | Pemba |
|---------------------------|--------|------|------|------|-------|-------|
| IndSilicoVA               | 0.64   | 0.73 | 0.55 | 0.65 | 0.67  | 0.42  |
| Bayesian Factor Model     | 0.79   | 0.82 | 0.66 | 0.75 | 0.78  | 0.57  |
| FARVA                     | 0.77   | 0.79 | 0.66 | 0.67 | 0.64  | 0.55  |
| LCVA-S: constant weights  | 0.75   | 0.68 | 0.68 | 0.7  | 0.75  | 0.61  |
| LCVA-S: new weights       | 0.74   | 0.67 | 0.67 | 0.7  | 0.74  | 0.59  |
| LCVA-M: domain-level mixture | 0.77  | 0.68 | 0.76 | 0.69 | 0.77  | 0.63  |
| LCVA-M: domain-cause-level mixture | 0.77 | 0.66 | 0.74 | 0.68 | 0.74  | 0.62  |

Figure 15: Individual-level top cause prediction accuracy (left) and CSMF accuracy (right) of different models. The proposed models are fitted with $K = 15$. The cells are colored by the relative difference from the InSilicoVA results, where blue cells indicate higher accuracy and red cells indicate lower accuracy. All variations of LCVA demonstrate good performance that is significantly higher than InSilicoVA except in one experiment. Bayesian factor model and FARVA also show comparable improvements.
C.3 Additional results on resampled PHMRC experiment

In this subsection, we first provide additional summaries for the relative improvement over the Bayesian factor model, in addition to the results presented in the main paper. The mean, and the 25th and 75th percentiles of the relative improvements of all models over the Bayesian factor model are summarized in Table 1. Except for the top cause accuracy in Dar es Salaam, the two versions of LCVA-M improves from Bayesian factor models in all other cases, and the improvements are quite dramatic in Mexico city, Andhra Pradesh, Uttar Pradesh, and Bohol. Figure 16 show the distribution of relative improvements over all 300 resampled datasets for each model. Across all experiments, the average improvements of LCVA-M are similar. Take LCVA-M with domain-cause-level mixture for example, we observe on average around 19% and 14% improvement in terms of the top cause accuracy and CSMF accuracy respectively.

![Figure 16: Distribution of relative improvement over the Bayesian factor model, i.e., \((Acc_m - Acc_{BF})/ACC_{BF}\), for each model \(m\) in terms of the top cause accuracy (left) and CSMF accuracy (right) by different models on all 300 resampled target datasets.](image)

In addition to the comparison with Bayesian factor model, Figure 17 show the relative improvements from InSilicoVA computed as \((Acc_m - Acc_{insa})\) for all models considered. We
| Site     | Model                  | Top Cause Accuracy | CSMF Accuracy |
|----------|------------------------|--------------------|---------------|
| Mexico   | FARVA                  | 47% (21%, 66%)     | -22% (-41%, -7%) |
|          | LCVA-S: constant weights | 44% (12%, 62%)     | 20% (9%, 31%) |
|          | LCVA-S: new weights    | 24% (9%, 38%)      | 15% (3%, 25%) |
|          | LCVA-M: domain-level mixture | 46% (19%, 55%)  | 20% (6%, 32%) |
|          | LCVA-M: domain-cause mixture | 53% (23%, 64%) | 23% (6%, 34%) |
| AP       | FARVA                  | 17% (4%, 22%)      | -13% (-30%, 0%) |
|          | LCVA-S: constant weights | 11% (-2%, 21%)     | 8% (-3%, 20%) |
|          | LCVA-S: new weights    | 4% (-10%, 10%)     | 4% (-8%, 18%) |
|          | LCVA-M: domain-level mixture | 12% (0%, 25%)  | 8% (-1%, 20%) |
|          | LCVA-M: domain-cause mixture | 13% (1%, 26%) | 8% (-3%, 23%) |
| UP       | FARVA                  | 22% (8%, 37%)      | 0% (-18%, 18%) |
|          | LCVA-S: constant weights | 11% (0%, 21%)     | 15% (3%, 25%) |
|          | LCVA-S: new weights    | 3% (-14%, 12%)     | 11% (-3%, 22%) |
|          | LCVA-M: domain-level mixture | 17% (4%, 26%)  | 19% (7%, 26%) |
|          | LCVA-M: domain-cause mixture | 16% (3%, 28%) | 19% (6%, 30%) |
| Dar      | FARVA                  | 3% (-9%, 19%)      | -24% (-39%, -11%) |
|          | LCVA-S: constant weights | 6% (-5%, 14%)     | 8% (-3%, 19%) |
|          | LCVA-S: new weights    | 3% (-10%, 16%)     | 13% (4%, 22%) |
|          | LCVA-M: domain-level mixture | -5% (-22%, 9%)  | 1% (-10%, 11%) |
|          | LCVA-M: domain-cause mixture | -5% (-19%, 8%) | 2% (-10%, 10%) |
| Bohol    | FARVA                  | 20% (-3%, 45%)     | -13% (-30%, 1%) |
|          | LCVA-S: constant weights | 29% (8%, 56%)     | 22% (6%, 38%) |
|          | LCVA-S: new weights    | 19% (2%, 35%)      | 20% (4%, 32%) |
|          | LCVA-M: domain-level mixture | 29% (-1%, 56%) | 24% (7%, 39%) |
|          | LCVA-M: domain-cause mixture | 30% (-5%, 59%) | 25% (6%, 39%) |
| Pemba    | FARVA                  | -20% (-36%, 4%)    | -22% (-42%, 1%) |
|          | LCVA-S: constant weights | 9% (-11%, 26%)     | 6% (-11%, 16%) |
|          | LCVA-S: new weights    | -11% (-42%, 13%)   | 1% (-25%, 15%) |
|          | LCVA-M: domain-level mixture | 4% (-26%, 25%) | 6% (-12%, 18%) |
|          | LCVA-M: domain-cause mixture | 7% (-18%, 25%) | 6% (-9%, 19%) |

Table 1: Mean relative improvement compared to Bayesian Factor model, i.e., \((\text{Acc}_m - \text{Acc}_{BF})/\text{ACC}_{BF}\), for for each model \(m\) in terms of the top cause accuracy and CSMF accuracy. The intervals are 25th and 75th percentiles of the improvement.
did not compute the percentage improvement because the accuracy metrics of InSilicoVA on many resampled datasets are too low, which makes the percentage improvement very high and not easily readable.

![Boxplot of absolute improvement over InSilicoVA in top cause accuracy (left) and CSMF accuracy (right) by different models on 50 resampled target domain datasets based on deaths from each of the six sites.](image)

Figure 17: Boxplot of absolute improvement over InSilicoVA (McCormick et al., 2016) in top cause accuracy (left) and CSMF accuracy (right) by different models on 50 resampled target domain datasets based on deaths from each of the six sites.

D Additional results when limited target data is available

In this section, we present additional results on the analysis of calibrating LCVA and InSilicoVA with additional labeled data from the target domain.
D.1 Cause of death aggregation

Table 2 shows the three levels of the cause list we use in the calibration process. The algorithms are fitted on the original 34 cause and the individual level predictions are aggregated to the other two broader categories for calibration. The result with the 34-cause list is presented in the main manuscript.

D.2 Calibration on the aggregated cause list

Figure 18 compares the calibrated CSMF accuracy on the 11-cause list. Except for Dar es Salaam, the joint modeling approach LCVA with partial labels achieve comparable or higher CSMF accuracy than the calibration approach. Figure 19 compares the calibrated CSMF accuracy on the broader 5-cause list, and the performance of the calibrated methods are similar. This analysis suggest that our approach to directly model the joint distribution of symptoms across domains is beneficial both with and without labeled data from the target domain, with the greatest improvements from the conditional independence model when the number of causes to consider is large.

E Convergence Analysis

In this section, we present additional results of the MCMC convergence analysis. We focus on the multi-domain LCVA model with domain-level shrinkage here. Other variants of the model show similar behavior in the posterior draws. Figure 20 to 25 shows the trace plots of the estimated CSMF vector and site similarity parameter $\eta$ from three independent MCMC chains in the prediction stage for each of the six sites in the experiment in Section 5.2 of the main paper. Each prediction MCMC chain is fitted with a separate training
| 34 Causes                                      | 11 Causes                                      | 5 Causes                                      |
|-----------------------------------------------|-----------------------------------------------|----------------------------------------------|
| Acute Myocardial Infarction                   | Disease of the Circulatory System             | Circulatory                                  |
| Other Cardiovascular Diseases                 | Disease of the Circulatory System             | Circulatory                                  |
| Stroke                                        | Disease of the Circulatory System             | Circulatory                                  |
| Bite of Venomous Animal                       | External                                      | External                                     |
| Drowning                                      | External                                      | External                                     |
| Falls                                         | External                                      | External                                     |
| Fires                                         | External                                      | External                                     |
| Homicide                                      | External                                      | External                                     |
| Other Injuries                                | External                                      | External                                     |
| Poisonings                                    | External                                      | External                                     |
| Road Traffic                                  | External                                      | External                                     |
| Suicide                                       | External                                      | External                                     |
| AIDS                                          | Infectious and Parasitic Diseases             | Infectious                                   |
| Diarrhea/Dysentery                            | Infectious and Parasitic Diseases             | Infectious                                   |
| Malaria                                       | Infectious and Parasitic Diseases             | Infectious                                   |
| Other Infectious Diseases                     | Infectious and Parasitic Diseases             | Infectious                                   |
| Pneumonia                                     | Infectious and Parasitic Diseases             | Infectious                                   |
| TB                                            | Infectious and Parasitic Diseases             | Infectious                                   |
| Maternal                                      | Maternal                                      | Maternal                                     |
| Cirrhosis                                     | Gastrointestinal disorders                    | Non-Communicable                             |
| Epilepsy                                      | Mental and Nervous System Disorders           | Non-Communicable                             |
| Breast Cancer                                 | Neoplasms                                     | Non-Communicable                             |
| Cervical Cancer                               | Neoplasms                                     | Non-Communicable                             |
| Colorectal Cancer                             | Neoplasms                                     | Non-Communicable                             |
| Esophageal Cancer                             | Neoplasms                                     | Non-Communicable                             |
| Leukemia/Lymphomas                            | Neoplasms                                     | Non-Communicable                             |
| Lung Cancer                                   | Neoplasms                                     | Non-Communicable                             |
| Prostate Cancer                               | Neoplasms                                     | Non-Communicable                             |
| Stomach Cancer                                | Neoplasms                                     | Non-Communicable                             |
| Diabetes                                      | Nutritional and Endocrine Disorders           | Non-Communicable                             |
| Other Non-communicable Diseases               | Other Noncommunicable Diseases                | Non-Communicable                             |
| Renal Failure                                 | Renal Disorders                               | Non-Communicable                             |
| Asthma                                        | Respiratory disorders                         | Non-Communicable                             |
| COPD                                          | Respiratory disorders                         | Non-Communicable                             |

Table 2: Two levels of aggregation of the 34 causes of death in the PHMRC dataset in the calibration analysis.

The stage where six training MCMC chains were fitted and stacked to estimate the posterior distribution of the latent class parameters. Overall, the chains mix well and converge to the same distribution across the three independent runs.
Figure 18: Comparison of LCVA-M with domain-level mixture and InSilicoVA when the cause of death of 30% of data in the target domain are known. The causes of deaths have been aggregated into 11 broad categories. The horizontal line indicates the CSMF accuracy of the two methods before calibrating to the labeled deaths. The box plot shows the distribution of the calibrated CSMF accuracy over 50 resampled labeled dataset.
Figure 19: Comparison of LCVA-M with domain-level mixture and InSilicoVA when the cause of death of 30% of data in the target domain are known. The causes of deaths have been aggregated into 5 broad categories. The horizontal line indicates the CSMF accuracy of the two methods before calibrating to the labeled deaths. The box plot shows the distribution of the calibrated CSMF accuracy over 50 resampled labeled dataset.
Figure 20: Mexico City as target site: Trace plots for each CSMF, $\pi_c$ (top), and site similarity parameter $\eta_g$ (bottom) from three chains. The burn-in period is shaded in gray.
Figure 21: Andhra Pradesh as target site: Trace plots for each CSMF, $\pi_c$ (top), and site similarity parameter $\eta_g$ (bottom) from three chains. The burn-in period is shaded in gray.
Figure 22: Uttar Pradesh as target site: Trace plots for each CSMF, $\pi_c$ (top), and site similarity parameter $\eta_g$ (bottom) from three chains. The burn-in period is shaded in gray.
Figure 23: Dar es Salaam as target site: Trace plots for each CSMF, $\pi_c$ (top), and site similarity parameter $\eta_g$ (bottom) from three chains. The burn-in period is shaded in gray.
Figure 24: Bohol as target site: Trace plots for each CSMF, $\pi_c$ (top), and site similarity parameter $\eta_g$ (bottom) from three chains. The burn-in period is shaded in gray.
Figure 25: Pemba Island as target site: Trace plots for each CSMF, $\pi_c$ (top), and site similarity parameter $\eta_g$ (bottom) from three chains. The burn-in period is shaded in gray.