Tree-informed Bayesian multi-source domain adaptation: cross-population probabilistic cause-of-death assignment using verbal autopsy

ZHENKE WU[*,1,2], ZEHANG R. LI[3], IRENA CHEN[1], MENGBING LI[1]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA*

[2]*Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI 48109, USA*

[3]*Department of Statistics, University of California, Santa Cruz, CA 95064, USA*

[*]zhenkewu@umich.edu

SUMMARY

Determining causes of deaths (COD) occurred outside of civil registration and vital statistics systems is challenging. A technique called verbal autopsy (VA) is widely adopted to gather information on deaths in practice. A VA consists of interviewing relatives of a deceased person about symptoms of the deceased in the period leading to the death, often resulting in multivariate binary responses. While statistical methods have been devised for estimating the cause-specific mortality fractions (CSMFs) for a study population, continued expansion of VA to new populations (or "domains") necessitates approaches that recognize between-domain differences while capitalizing on potential similarities. In this paper, we propose such a domain-adaptive method that integrates external between-domain similarity information encoded by a pre-specified rooted weighted tree. Given a cause, we use latent class models to characterize the conditional distributions of the responses that may vary by domain. We specify a logistic stick-breaking Gaussian diffusion process prior along the tree for class mixing weights with node-specific spike-and-slab priors to pool information between the domains in a data-driven way. Posterior inference is conducted via a scalable variational Bayes algorithm. Simulation studies show that the domain adaptation enabled by the proposed method improves CSMF estimation and individual COD assignment. We also illustrate and evaluate the method using a validation data set. The paper concludes with a discussion on limitations and future directions.

Key words: Domain Adaptation; Latent Class Models; Spike-and-Slab Prior; Variational Bayes; Verbal Autopsy

[*]To whom correspondence should be addressed.

1. INTRODUCTION

1.1 *Verbal Autopsy (VA): background*

Patterns of mortality and causes of death at the community level are critical to informing public health policies, tracking trends, and prioritizing interventions for local governments and public health officials. Civil registration systems that track births, deaths and their causes provide the basis for countries to identify their most pressing health issues. Despite ongoing global efforts to strengthen the civil registration and vital statistics (CRVS) system, two-thirds of 56 million annual deaths go unrecorded, especially in low- and middle-income countries (LMIC), leaving glaring gaps for reliable mortality information (World Health Organization, 2021). Verbal autopsy (VA) is one of the most well-established and realistic methods to collect information about cause of death (COD) in these situations when medically certified cause of death is unavailable. VAs collect information on deaths by interviewing caregivers (or individuals who witnessed the death) of the deceased. Typically, information about healthcare access, demographic information, and various indicators of symptoms leading to the death are collected. See Chandramohan *and others* (2021) for a recent review of historical developments, ongoing efforts to standardize VA instruments and implications for LMIC.

The central analytic goal is to use VA data to derive population-level cause-specific mortality fractions (CSMFs) and to produce individual-level COD assignment. In particular, VA data contain pertinent information on signs, symptoms, and circumstances leading to death, generically referred to as "symptoms" in this paper. These symptoms are often coded into binary "Yes"/"No" answers, resulting in data sets that contain multiple binary responses for each death. Algorithmic and probablistic methods have been developed to automate the task of estimating CSMFs and individual-level COD assignment. See Li *and others* (2021*b*) for an excellent comprehensive review of the major methods and software as well as the references therein for recent methodological improvements. Developments in analytic methods and reproducible open-source software for VA have greatly fostered confidence in large-scale implementations of VA in many LMICs.

1.1.1 *Statistical challenges in domain adaptation in VA research*   However, one emerging analytical challenge in expanding VA to new populations is in developing approaches that recognize and address potential differences between the existing and new populations in terms of the joint distribution of causes of death and VA responses. Such a problem is an example of "domain adaptation" in machine learning literature

(e.g., Pan and Yang, 2009), but has received relatively little attention in VA research. In particular, CSMFs comprise a vector of population-level marginal probabilities of the causes and may differ by domain; this is most natural because a cause may differentially contribute to deaths occurred in different study populations. The conditional distribution of the VA responses given a cause characterizes symptom-cause relationships and may also differ by domain; we focus our paper and model formulation on addressing this aspect of domain adaptation.

It is well known that more accurate estimation of the conditional distribution of the multivariate binary VA responses given a cause can result in substantial improvements to CSMF estimation performance (e.g., Kunihama *and others*, 2020; Li, McComick and Clark, 2020). To acknowledge potential between-domain differences in these conditional distributions given a cause, it is therefore tempting to directly estimate them for each domain separately. However, in a domain with few sampled deaths due to a cause, such direct estimates are often vulnerable to statistical instability, hindering accurate CSMF estimation. This issue worsens still if for that same cause the numbers of sampled deaths are small in multiple domains. In such cases, pooling information from similar domains would improve the estimation of these conditional distributions which in turn would propagate to improving the estimation of CSMFs. On the other hand, an extreme complete-pooling approach that forces domains to have an identical conditional distribution of the VA symptoms given any cause is restrictive and would obscure the study of important between-domain variations in response patterns (e.g., King and Lu, 2008; McCormick *and others*, 2016). Data-driven pooling of information between the domains for each cause is desirable.

1.1.2 *Existing literature* Here we briefly describe a few existing work related to domain adaptation in verbal autopsy studies and how the proposed method differ from them. Datta *and others* (2021) and Fiksel *and others* (2021) developed methods that calibrate CSMF estimates obtained from VA algorithms trained on a training data set to produce CSMF estimates in a new population. These calibration methods differ from our work in three important ways. First, such calibration methods only consider the estimated CSMFs from a list of trained VA algorithms, and are hence not designed for using individual-level information in the training data set to perform calibration. Second, the calibration relies on a small number of deaths with medically-confirmed causes in the new population. Third, causes often need to be manually combined before calibration can be applied to produce stable and meaningful results. Our work focus on using all individual-

level data from multiple populations (referred to as "source domains") with known causes of death, and cause-of-death assignment in a new population (referred to as "target domain"). The proposed method does not require known cause-of-death labels in the target domain or any ad hoc collapsing of causes.

Among a few methods that directly model the individual-level VA data under domain adaptation, one related work is Moran *and others* (2021) that introduces a factor regression method to let the conditional distribution of the VA symptoms given a cause vary by additional individual-level covariates, which may include dummy domain indicators. This approach again is not designed for the scenario where cause-of-death labels are fully unobserved in the target domain. Our work is most related to Li *and others* (2021*c*), where a latent class model framework was proposed to model the conditional dependence relationship among symptoms with improved interpretability and computational speed. However VAs collected from different domains are treated as marginally independent data sets. Our work significantly extends Li *and others* (2021*c*) by proposing a framework that can integrate additional external structural information across the domains. This allows us to efficiently pool information across domains and improve the stability of the estimates when some causes are rarely observed.

1.2 *Main contributions*

This paper develops a novel tree-integrative framework for CSMF estimation and individual-level COD assignment based on latent class models (Lazarsfeld, 1950) that jointly models multivariate binary data obtained from multiple source domains and a target domain. Our framework explicitly acknowledges domain-by-domain variation in the distribution of causes and distribution of symptoms given causes. Most importantly, it takes into account the structural similarities among deaths from related domains. Our main methodological contributions are two folds.

First, we propose a data-driven pooling of information across domains via a pre-specified hierarchy represented by a rooted weighted tree. This approach is shown to encourage similar conditional dependence structure across domains while recognizing important between-domain differences resulting in more accurate CSMF estimation. Simulation studies show the proposed approach has better performance compared to estimates that either completely, incorrectly, or minimally pool information across domains. Although the method is general, in this paper, we illustrate the method by a domain tree defined by geographical region of

each study site, which serves as a proxy for potential regional variations in factors which may drive differences in the conditional distributions of symptoms given a cause, e.g., VA interviewer training, culture in symptom disclosure of a deceased, etc. As a secondary feature, the proposed approach also uses a hierarchy over the causes to enable information pooling between causes so that a rare cause can be pooled with similar causes to produce more stable estimates of the class-specific response profiles.

Second, we propose a tree-based measure of dissimilarity in symptom-cause relationship between the target domain and each of the source domains, separately for each cause. The proposed measure admits rich interpretation of empirical evidence about the manner in which causes differ in between-domain similarities. For example, causes with highly recognizable and specific symptoms (e.g., "Drowning") may have symptom-cause conditional distributions that remain similar regardless of the domains, while less so for other causes with complex etiologies that are prone to differential reporting patterns across the domains.

Paper organization   The rest of the paper is organized as follows. Section 2 reviews tree-related terminologies and presents the proposed model. Section 3.1 specifies prior distributions. A variational Bayes algorithm is presented in Section 4. Section 5 conducts simulation studies to illustrate the operating characteristics of the proposed method. In Section 6, we use a validation data set to illustrate the method. The paper concludes with a brief summary and discussion on limitations and some future directions.

2. Model

We first introduce necessary terminologies and notations for characterizing a rooted weighted tree. The proposed nested latent class model (NLCM) is then formulated for deaths, each with an observed link to a leaf in a tree over source and target domains.

2.1   *Rooted weighted trees*

A rooted tree is a graph $\mathcal{T} = (\mathcal{V}, E)$ with node set $\mathcal{V}$ and edge set E where there is a root u_0 and each node has at most one parent node. Let $p = |\mathcal{V}|$ represent the total number of leaf and non-leaf nodes. Let $\mathcal{V}_{\mathsf{leaf}} \subset \mathcal{V}$ be the set of leaves (i.e., nodes without children), and $p_{\mathsf{leaf}} = |\mathcal{V}_{\mathsf{leaf}}| < p$. We typically use u to denote any node ($u \in \mathcal{V}$) and v to denote any leaf ($v \in \mathcal{V}_{\mathsf{leaf}}$). Each edge in a rooted tree defines a *clade*: the group of

leaves below it. Splitting the tree at an edge creates a partition of the leaves into two groups. For any node $u \in \mathcal{V}$, the following notations apply: $c(u)$ is the set of offspring of u, $pa(u)$ is the parent of u, $d(u)$ is the set of descendants of u including u, and $a(u)$ is the set of ancestors of u including u. At the top of Figure 1, a hypothetical tree for $G = 4$ source domains and one target domain with $p = 8$ and $p_{\text{leaf}} = 5$ is shown. If $u = 2$, then $c(u) = \{5, 6\}$, $pa(u) = \{1\}$, $d(u) = \{2, 5, 6\}$, and $a(u) = \{1, 2\}$. See Figure 3 (top margin) for an instance of a nested hierarchy for six domains where VA data are collected: $p_{\text{leaf}} = 6$ leaves representing six study sites, and $p - p_{\text{leaf}} = 3$ non-leaf nodes subsuming the six leaf descendants (root node representing "global"; two internal nodes representing two countries, "India" and "Tanzania").

Edge-weighted graphs appear as a model for numerous problems where nodes are linked with edges of different weights. In particular, the edges in $\mathcal{T}$ are attached with weights where $w : E \rightarrow \mathbb{R}^+$ is a weight function. Let $\mathcal{T}_w = (\mathcal{T}, w)$ be a rooted weighted tree. A path in a graph is a sequence of edges which joins a sequence of distinct vertices. For a path P in the tree connecting two nodes, $w(P)$ is defined as the sum of all the edge weights along the path, often referred to as the "length" of P. The distance between two vertices u and u', denoted by $dist_{\mathcal{T}_w}(u, u')$ is the length of a shortest (with minimum length) (u, u')-path. $dist_{\mathcal{T}_w}$ is a distance: it is symmetric and satisfies the triangle inequality. In this paper, we use w_u to represent the edge length between a node u and its parent node $pa(u)$. w_u is fully determined by $\mathcal{T}_w$. For the root u_0, there are no parents, i.e. $pa(u_0) = \emptyset$; we set $w_{u_0} = 1$. In VA contexts, although w_u may be specified via the dendrogram resulting from a hierarchical clustering of domain-level covariates, in Section 6 we will set $w_u = 1$ to use minimal external domain similarity information (geographical region) for simpler exposition.

2.2   *Nested latent class models (Nested LCM)*

Although LCMs work for multiple discrete responses of more than two levels (e.g., Lazarsfeld, 1950), in this paper, we present the model for multivariate binary responses for simpler exposition.

Notations   Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iJ})^\mathsf{T} \in \{0, 1\}^J$ be a vector of J binary responses for subject $i \in [N]$ where N is the total number of subjects; here $[Q] = \{1, \ldots, Q\}$ generically represents positive integers no greater than a positive integer Q. Let (Y_i, D_i) represent (cause of death, domain), where Y_i takes its value from $\{1, \ldots, C\}$ indicating the cause of death among a total of C pre-specified causes; let $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^\mathsf{T}$. D_i takes its value from $\{0, 1, \ldots, G\}$ indicating subject i's domain membership: 0 for target domain, and 1 to G for G

pre-specified source domains. Let $\boldsymbol{D} = (D_1, \ldots, D_N)^{\mathsf{T}}$. Throughout the paper, D_i is assumed to be observed for all subjects; Y_i is assumed to be observed for subjects in the source domain $\{i : D_i \neq 0\}$ but unobserved for subjects in the target domain $\{i : D_i = 0\}$. Let $\boldsymbol{Y}^{\mathsf{obs}} = \{Y_i : D_i \neq 0\}$ and $\boldsymbol{Y}^{\mathsf{mis}} = \{Y_i : D_i = 0\}$; we then have $\boldsymbol{Y} = (\boldsymbol{Y}^{\mathsf{obs}}, \boldsymbol{Y}^{\mathsf{mis}})^{\mathsf{T}}$. Let $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)^{\mathsf{T}}$ be an $N \times J$ binary data matrix for all subjects. $\boldsymbol{D}$ maps every row of data $\mathbf{X}$ to a leaf in the tree for domains $\mathcal{T}_w$. Similarities between domains are then characterized by between-domain distances in $\mathcal{T}_w$. Finally, let $\mathcal{D} = (\mathbf{X}, \boldsymbol{Y}^{\mathsf{obs}}, \boldsymbol{D})$ represent the data from all the domains. We use "pr$(A \mid B)$" to represent the conditional density of random variable(s) in A given B.

2.2.1 *Model formulation* We assume the following model specifications for $\mathcal{D}$:

$$\text{cause of death}: \ Y_i \mid D_i = g \sim \mathsf{Categorical}_C(\boldsymbol{\pi}^{(g)}), \tag{2.1}$$

$$\text{latent class}: \ Z_i \mid Y_i = c, D_i = g \sim \mathsf{Categorical}_K(\boldsymbol{\lambda}^{(c,g)}), \tag{2.2}$$

$$\text{responses}: \ X_{ij} \mid Z_i = k, Y_i = c \overset{\mathsf{indep.}}{\sim} \mathsf{Bernoulli}(\theta_{jk}^{(c)}), j \in [J] \tag{2.3}$$

for $i \in [N]$, $g \in \{0\} \cup [G]$, where the population parameters $\boldsymbol{\pi}^{(g)} = (\pi_1^{(g)}, \ldots, \pi_c^{(g)})^{\mathsf{T}}$ with $\sum_{c=1}^{C} \pi_c^{(g)} = 1$ are referred to as "cause-specific mortality fractions" (CSMFs). Importantly, $\{\boldsymbol{\pi}^{(g)}, g = 0, 1, \ldots, G\}$ are not constrained to be identical. We seek to estimate $\boldsymbol{\pi}^{(0)}$ and $\{Y_i : D_i = 0\}$.

On the latent classes, $\boldsymbol{Z} = (Z_1, \ldots, Z_N)^{\mathsf{T}} \in [K]^N$ is a vector of class memberships for all subjects; $\boldsymbol{\lambda}^{(c,g)} = (\lambda_1^{(c,g)}, \ldots, \lambda_K^{(c,g)})^{\mathsf{T}}$ is a vector of class weights that sum to one: $\sum_{k=1}^{K} \lambda_k^{(c,g)} = 1$. See Remark 2.1 for the nuisance role of $\boldsymbol{Z}$ and the nuisance notion of "class" that are introduced for inducing conditional stochastic dependence among VA responses given any pair of cause and domain. On terminology, by Equation (2.2), the K latent classes that Z_i can take are nested within a cause of death c; we refer to the model as "nested" latent class model.

On the class-specific response probabilities, for a subject died of cause c, $\theta_{jk}^{(c)} \in [0, 1]$ is the positive response probability for item j in class k. Let $\boldsymbol{\theta}_{\cdot k}^{(c)} = (\theta_{1k}^{(c)}, \ldots, \theta_{Jk}^{(c)})^{\mathsf{T}}$ be the vector of the k-th class response probability profile; let $\boldsymbol{\Theta}^{(c)} = \left(\boldsymbol{\theta}_{\cdot 1}^{(c)}, \ldots, \boldsymbol{\theta}_{\cdot K}^{(c)}\right)$ collect these probabilities into a $J \times K$ matrix with (j, k)-th element $\theta_{jk}^{(c)}$. Note that $\boldsymbol{\Theta}^{(c)}$ may vary by cause c. This admits flexible characterization of symptoms that may have distinct distributions for different true causes-of-death. In addition, for any given c, $\boldsymbol{\Theta}^{(c)}$ is assumed domain-invariant (source or target) to facilitate shared interpretations. Conversely, letting the class response profiles vary greatly between domains would weaken the diagnostic explanability of the VA questionnaire

items. See Figure 1 for a schematic representation of the data generating process under the proposed model.

For any given cause c, despite shared $\boldsymbol{\Theta}^{(c)}$ across the domains, $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = g)$ may differ by domain g as a result of distinct class weights $\boldsymbol{\lambda}^{(c,g)}$ across the domains. This is readily seen from Equations (2.2) and (2.3) which imply that the conditional distributions are fully parameterized by $(\boldsymbol{\Theta}^{(c)}, \boldsymbol{\lambda}^{(c,g)})$:

$$\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = g) = \sum_{k=1}^{K} \lambda_k^{(c,g)} \cdot \prod_{j=1}^{J} \left\{ \theta_{jk}^{(c)} \right\}^{X_{ij}} \left\{ 1 - \theta_{jk}^{(c)} \right\}^{1-X_{ij}}, g = 0, 1, \ldots, G. \tag{2.4}$$

If $\boldsymbol{\lambda}^{c,g} = \boldsymbol{\lambda}^{(c,g')}$ for any $g, g' = 0, 1, \ldots, G$, Equation (2.4) simplifies to $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c)$.

REMARK 2.1 The model treats $\boldsymbol{Z}$ and the notion of "class" as technical nuisances that are introduced for the sole purpose of flexibly modeling $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i, D_i)$ (Dunson and Xing, 2009). In particular, when $K \geqslant 2$ and $\{\boldsymbol{\Theta}^{(c)}\}$ differ by c, although Equation (2.3) assumes conditional independence given a latent class and a cause, by integrating over Z_i with probabilities $\boldsymbol{\lambda}^{(Y_i,D_i)}$, we induce stochastic dependence among the J components of $\boldsymbol{X}_i$ (Equation (2.4)). Setting $K = 1$ would assume that the VA responses are mutually independent given any pair of cause and domain.

REMARK 2.2 Equations (2.1) to (2.3) under $D_i = 0$ are equivalent to a $K \cdot C$-class LCM for $\{\boldsymbol{X}_i : D_i = 0\}$ parameterized by $(\boldsymbol{\pi}^{(0)}, \boldsymbol{\lambda}^{(c,0)}, \boldsymbol{\Theta}^{(c)}, c \in [C])$. Fortunately, based on data from the source domains, the multivariate binary VA response data $\{\boldsymbol{X}_i : D_i \neq 0\}$ tabulated by the observed causes of deaths ($\boldsymbol{Y}^{\mathsf{obs}}$) provide direct information for estimating $\boldsymbol{\Theta}^{(c)}$ that is shared across the domains.

3. PRIORS

3.1 *Tree-structured shrinkage prior*

3.1.1 *Motivation and overview* Estimating target domain CSMFs and individual cause-of-death assignment rely on efficient learning of $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = 0)$, the joint distribution of multivariate binary VA responses given each cause. In particular, by Bayes rule

$$\mathbb{P}(Y_i = c \mid \boldsymbol{X}_i, D_i = g) = \frac{\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = g)\pi_c^{(g)}}{\sum_{c'=1}^{C} \mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c', D_i = g)\pi_{c'}^{(g)}}, g = 0, 1, \ldots, G. \tag{3.5}$$

When $g = 0$, even with known $\boldsymbol{\pi}^{(0)}$, a poor estimate of $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = 0)$ can adversely impact the cause-of-death assignment on an individual level; when $\boldsymbol{\pi}^{(0)}$ is unknown as in our context, a good estimate of the conditional distribution $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = 0)$ remains critical to stable statistical estimation, e.g.,

via expectation-maximization for finite mixture models. The same issue persists for accurate individual-level cause-of-death assignment when $g \neq 0$ for which $\boldsymbol{\pi}^{(g)}$ can be directly estimated. However, obtaining a good estimate of $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = g)$ is often challenging when there exist small or even zero cell counts for particular combinations of (c, g), which renders direct estimation of $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = g)$ statistically unstable if not impossible.

This motivates us to take advantage of potential between-domain similarities. We achieve this aim by learning $G + 1$ conditional distributions $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = g), g = 0, 1, \ldots, G$, in a data-driven way for causes $c = 1, \ldots, C$, respectively. Consider any cause c, by Equation (2.4), $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = g)$ is fully parameterized by $(\boldsymbol{\Theta}^{(c)}, \boldsymbol{\lambda}^{(c,g)})$. Because $\boldsymbol{\Theta}^{(c)}$ is shared across domains, the $G + 1$ vectors of class-mixing weights $\{\boldsymbol{\lambda}^{(c,g)}, g = 0, 1, \ldots, G\}$ fully determines the between-domain differences in Equation (2.4). This points us to the strategy of encouraging *a priori* similar values of vectors $\boldsymbol{\lambda}^{(c,g)}$ across domain $g = 0, 1, 2, \ldots, G$; the degree to which they are similar may differ by cause. Data from all the domains can then be used to learn the degrees of optimal pooling between the domains for the C causes, respectively.

To achieve this aim, in Section 3.1.2, we introduce a tree-structured shrinkage prior for the $G + 1$ vectors of class-mixing weights $\{\boldsymbol{\lambda}^{(c,g)}, g = 0, 1, \ldots, G\}$ for each c. The prior is based on a logistic stick-breaking Gaussian process, diffused along a pre-specified rooted weighted domain tree with $G + 1$ leaves that encodes external between-domain similarity information. Also see Appendix E in the Supplementary Materials for a review of the general statistical strategy of specifying tree-structured shrinkage priors (Thomas *and others*, 2020; Li *and others*, 2021*a*) .

3.1.2 **Logistic stick-breaking Gaussian diffusion prior for $\boldsymbol{\lambda}^{(c,g)}$: integrating domain hierarchy**

Recall that $\mathcal{T}_w = (\mathcal{T} = (\mathcal{V}, E), w)$ represents a rooted weighted tree over domains, where the $G + 1$ leaves $\mathcal{V}_{\mathsf{leaf}}$ comprise G source domains and one target domain; see Section 5.2 for an example in the context of VA. Each domain comprises multiple independent observations $\{(\boldsymbol{X}_i, Y_i) : D_i = g\}$. Each domain g is one-to-one mapped to a leaf in $\mathcal{T}_w$. We specify a prior based on a logistic stick-breaking Gaussian process diffused along $\mathcal{T}_w$ and end at $G + 1$ leaves, inducing a prior distribution over the class weights $\boldsymbol{\lambda}^{(c,g)}, g = 0, 1, \ldots, G$. We first reparameterize $\boldsymbol{\lambda}^{(c,g)}$ with a stick-breaking representation: $\lambda_k^{(c,g)} = V_k^{(c,g)} \prod_{s<k}(1 - V_s^{(c,g)})$, for $k \in [K]$, where $0 < V_k^{(c,g)} < 1$, for $k \in [K-1]$ and $V_K^{(c,g)} = 1$. In particular, let $\eta_k^{(c,g)} = \sigma^{-1}(V_k^{(c,g)})$, $k \in [K-1]$, $g \in \mathcal{V}_{\mathsf{leaf}}$, where $\sigma(x) = 1/\{1 + \exp(-x)\}$ is the sigmoid function. The logistic stick-breaking parameterization

is completed by

$$\lambda_k^{(c,g)} = \{\sigma(\eta_k^{(c,g)})\}^{\mathbf{1}\{k<K\}} \prod_{s<k} \sigma(-\eta_s^{(c,g)}), k \in [K], \tag{3.6}$$

where $\mathbf{1}\{A\}$ is indicator function and equals 1 if statement A is true and 0 otherwise. This reparametrization

lends itself to simple and accurate posterior inference via variational Bayes algorithms.

For a leaf $g \in \mathcal{V}_{\mathsf{leaf}}$, let

$$\eta_k^{(c,g)} = \sum_{u \in a(g)} \xi_k^{(c,u)}, k \in [K-1]. \tag{3.7}$$

Note that $\eta_k^{(c,g)}$ is defined for leaves $g = \{0\} \cup [G]$ only and $\xi_{uk}^{(c,g)}$ is defined for all the nodes $u \in \mathcal{V}$. Finally,

for $c = 1, \ldots, C$, we specify

$$\xi_k^{(c,u)} = s_{cu}\alpha_k^{(c,u)}, \forall\ u \in \mathcal{V}, \tag{3.8}$$

$$\alpha_k^{(c,u)} \sim N(0, \tau_{\ell_u} w_u), \text{ independently for } k \in [K-1], \forall\ u \in \mathcal{V}, \tag{3.9}$$

$$s_{cu_0} = 1, \text{ and } s_{cu} \sim \mathsf{Bernoulli}(\rho_{c\ell_u}), \text{ independently for } u \in \mathcal{V} \setminus u_0, \tag{3.10}$$

$$\rho_{c\ell} \sim \mathsf{Beta}(a_{c\ell}, b_{c\ell}), \text{ independently for } \ell \in [L], \tag{3.11}$$

where $N(m', s')$ represents a Gaussian with mean m' and variance s'. In addition, $\ell_u \in [L]$ and maps node

$u \in \mathcal{V}$ (leaf or non-leaf) to one of L levels; this enables distinct degrees of diffusion for L non-overlapping

blocks of a pre-specified partition of the nodes. Let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_L)^{\mathsf{T}}$ be the diffusion variances for L levels

of nodes in the domain tree. Let $\boldsymbol{s} = \{s_{cu}, c \in [C], u \in \mathcal{V}\}$ be a $C \times p$ matrix of slab component indicators.

Let $\boldsymbol{\varrho}$ be a $C \times L$ matrix with (c, ℓ)-th element $\rho_{c\ell}$. Note that the spike-and-slab indicator probabilities $\rho_{c\ell_u}$

for a node u in the domain tree may vary by cause; relative to a more restrictive form $\rho_{c\ell_u} = \rho_{\ell_u}$, the present

specification has the additional flexibility of cause-specific degree of pooling between domains. When $K > 1$

and $s_{cu} = 1, \forall u \in \mathcal{V}_{\mathsf{leaf}}$, we would assume domains differ in class weights with probability one.

3.2   *Prior for other parameters*

We assume independent Dirichlet priors for the CSMFs in the source and target domains:

$$\boldsymbol{\pi}^{(g)} \stackrel{d}{\sim} \mathsf{Dirichlet}(\boldsymbol{d}^{(g)}), g = 0, 1, \ldots, G, \tag{3.12}$$

where $\boldsymbol{d}^{(g)} = (d_1^{(g)}, \ldots, d_C^{(g)})^{\mathsf{T}}$ is a vector of hyperparameters; let $\boldsymbol{d} = \{\boldsymbol{d}^{(g)} : g = 0, 1, \ldots, G\}$. In our

simulations, we simply use $\boldsymbol{d}^{(g)} = \mathbf{1}$ to represent a uniform prior over all cause which works well empirically.

In practice, informative knowledge can be incorporated by modifying these $\boldsymbol{d}^{(g)}$ hyperparameters to match prior numbers of observed deaths of each cause in domain g.

3.2.1 *Secondary feature of the method: integrating cause hierarchy* In a particular analysis, the number of causes can be large. Causes considered to be similar in nature and etiology would produce a symptom with similar probabilities. In addition, the number of deaths due to a cause can be small, resulting in unstable estimation of the response profiles if done separately from other causes. To overcome these issues, $\{\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(C)}\}$ is also equipped with a tree-structured prior with a pre-specified cause tree of C leaves that encodes between-cause similarities. For example, Figure 3 shows a cause tree in our VA application (left margin) representing a hierarchy of $p^*_{\text{leaf}} = 35$ causes and $p^* - p^*_{\text{leaf}} = 7$ internal nodes that represent coarser aggregated causes; we specify all edge weights to be one. By doing so, we encourage *a priori* similar values of $\boldsymbol{\Theta}^{(c)}$ across causes $c = 1, \ldots, C$. This facilitates optimal pooling of information over causes and overcomes statistical stability issues for rare causes that would otherwise require ad hoc manual cause aggregations (Datta *and others*, 2021). Although the proposed approach can accommodate two hierarchies (domain and cause), because domain adaption is our primary goal, in the following we will focus empirical evaluations on the use of domain hierarchy.

Let $\mathcal{T}^*_{w^*} = (\mathcal{T}^* = (\mathcal{V}^*, E^*), w^*)$ represent a rooted weighted tree, where the leaf set $\mathcal{V}^*_{\text{leaf}}$ represents distinct causes of death labeled as $c = 1, \ldots, C$. Each cause is mapped to one and only one leaf in $\mathcal{T}^*_{w^*}$. Note that $\boldsymbol{Y}$ maps every row of data $\mathbf{X}$ to a leaf in the tree for causes $\mathcal{T}^*_{w^*}$; but this link is only observed for deaths occurred in the source domains $\{i : D_i \neq 0\}$. Similarities between causes are then characterized by between-cause distances in $\mathcal{T}^*_{w^*}$. We then specify a logistic Gaussian diffusion prior:

$$\theta^{(c)}_{jk} = \text{expit}\left(\beta^{(c)}_{jk}\right), \quad \beta^{(c)}_{jk} = \sum_{u \in a(c)} \gamma^{(u)}_{jk}, c \in [C], \tag{3.13}$$

with Gaussian increments over the edges leading to each leaf:

$$\gamma^{(u)}_{jk} \sim N(0, \tau^*_{\ell^*_u} w^*_u), \tag{3.14}$$

where ℓ^*_u maps node u in the cause tree to one of L^* level; this is to allow distinct diffusion variances for L^* nonoverlapping blocks of a pre-specified partition of the nodes in the cause tree. Let $\boldsymbol{\tau}^* = (\tau^*_1, \ldots, \tau^*_{L^*})^\mathsf{T}$ be the vector of diffusion variances for the L^* sets of nodes in the cause tree. Unlike in Section 3.1.2, we choose not to use node-specific spike-and-slab priors in the cause tree, which is equivalent to less aggressive

shrinkage between causes and performs well in our simulation and validation studies.

Remark 3.1   Taken together, Sections 3.1.2 and 3.2.1 propose a prior for conditional distributions of the VA responses given any cause is a two-way tree-structured priors: i) the shrinkage among the domains in the columns is guided by a domain tree, and, ii) the shrinkage among the causes in the rows is guided by a cause tree. In particular, the shrinkage across causes is not domain-specific, but rather determined by information pooled across all domains. However, the shrinkage across domains is determined by a global-local structure, where we use τ_{ℓ_u} as diffusion variance parameter for all causes ("global"; τ_{ℓ_u} not indexed by c) and we use $\rho_{c\ell_u}$ to introduce cause-specific ("local"; $\rho_{c\ell_u}$ indexed by c) shrinkage of a leaf towards its parent ($\rho_{c\ell_u}$ closer to 0 or 1 for stronger or weaker shrinkage).

3.3   *Joint distribution*

The joint distribution is fully specified by Equations (2.1-2.3), (3.6-3.11), (3.12), and (3.13-3.14). We collect the unobserved quantities by $\mathbf{\Gamma} := \{\mathbf{Y}^{\mathsf{mis}}, \boldsymbol{\pi}^{(g)}, \mathbf{Z}, \boldsymbol{\alpha}^{(c,u)}, \gamma_{jk}^{(c)}, \boldsymbol{s}, \boldsymbol{\varrho}, c \in [C], g \in \{0\} \cup [G], j \in [J], k \in [K]\}$. We have the joint distribution of data $\mathcal{D}$ and $\mathbf{\Gamma}$ as follows:

$$
\begin{aligned}
\mathrm{pr}(\mathcal{D}, & \mathbf{\Gamma} \mid \mathcal{T}_w, \mathcal{T}_{w^*}^*, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\tau}, \boldsymbol{\tau}^*, \boldsymbol{d}) \\
&= \prod_{i=1}^{N} \prod_{g=0}^{G} \prod_{c=1}^{C} \left(\pi_c^{(g)} \prod_{k=1}^{K} \left\{ \lambda_k^{(c,g)} \prod_{j=1}^{J} \sigma(X_{ij}^* \beta_{jk}^{(c)}) \right\} \right)^{\mathbf{1}\{Z_i=k, Y_i=c, D_i=g\}} \\
&\times \prod_{c=1}^{C} \prod_{u \in \mathcal{V}} \prod_{k=1}^{K-1} \frac{1}{\sqrt{2\pi\tau_{\ell_u} w_u}} \exp\left(-\frac{1}{2\tau_{\ell_u} w_u} \left\{ \alpha_k^{(c,u)} \right\}^2 \right) \\
&\times \prod_{u \in \mathcal{V}^*} \prod_{j=1}^{J} \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\tau_{\ell_u^*}^* w_u^*}} \exp\left(-\frac{1}{2\tau_{\ell_u^*}^* w_u^*} \left\{ \gamma_{jk}^{(u)} \right\}^2 \right) \\
&\times \prod_{c=1}^{C} \prod_{u \in \mathcal{V}} \rho_{c\ell_u}^{s_{cu}} (1-\rho_{c\ell_u})^{1-s_{cu}} \cdot \prod_{g=0}^{G} \mathsf{Dirichlet}(\boldsymbol{\pi}^{(g)}; \boldsymbol{d}^{(g)}) \times \prod_{c=1}^{C} \prod_{\ell=1}^{L} \mathsf{Beta}(\rho_{c\ell}; a_{c\ell}, b_{c\ell}),
\end{aligned}
\tag{3.15}
$$

where Bernoulli likelihood components $\sigma(X_{ij}^* \beta_{jk}^{(c)}) := \{\theta_{jk}^{(c)}\}^{X_{ij}} \{1 - \theta_{jk}^{(c)}\}^{1-X_{ij}}$ with $X_{ij}^* := 2X_{ij} - 1$. Our primary quantity of interest is the CSMFs in the target domain $\boldsymbol{\pi}^{(0)}$ and the individual-specific cause-specific posterior probabilities $\mathbb{P}(Y_i = c \mid D_i = 0, \mathcal{D})$, $c \in [C]$.

The directed acyclic graph (DAG) in Appendix Figure 2 in the Supplementary Materials shows the relationship between the observables, unknown quantities and hyper-parameters.

4. Bayesian inference algorithms

Calculating a posterior distribution often involves intractable high-dimensional integration over the unknowns in the model. Traditional sequential sampling approaches such as Markov chain Monte Carlo (MCMC) remains a widely used inferential tool based on approximate samples from the posterior distribution. They can be powerful in evaluating multidimensional integrals. However, they do not guarantee closed-form posterior distributions. Variational inference (VI) is a popular alternative to MCMC for approximating the posterior distribution and has been widely used in machine learning and gaining interest in statistics (e.g., Blei, Kucukelbir and McAuliffe, 2017; Ormerod and Wand, 2010). In particular, VI has also been used for fitting the classical LCMs (e.g., Grimmer, 2011). VI requires a user-specified family of distributions that can be expressed in tractable forms while being flexible enough to approximate the true posterior; the approximating distributions and their parameters are referred to as "variational distributions" and "variational parameters", respectively. VI algorithms find the best variational distribution that minimizes the Kullback-Leibler (KL) distance between the variational family and the true posterior distribution. VI has been widely applied in Gaussian (Carbonetto, Stephens *and others*, 2012; Titsias and Lázaro-Gredilla, 2011) and binary likelihoods (e.g., Jaakkola and Jordan, 2000; Thomas *and others*, 2020). Also see Blei, Kucukelbir and McAuliffe (2017) for a detailed review. We use VI because it is fast, bypasses infeasible analytic integration or data augmentation that is otherwise needed for MCMC under Dirac spike components and prior-likelihood non-conjugacy (Tüchler, 2008), and enables data-driven selection of hyperparameters via approximate empirical Bayes (Step 3, Appendix A in the Supplementary Materials).

We wish to obtain the marginal posterior distributions $\mathrm{pr}(\boldsymbol{\pi}^{(0)} \mid \mathcal{D})$ and $\mathrm{pr}(Z_i \mid \mathcal{D})$. We conduct posterior inference via variational inference. We assume the variational distributions can factorize as follows:

$$q(\boldsymbol{\Gamma}) = \prod_{g=0}^{G} q(\boldsymbol{\pi}^{(g)}) \prod_{c=1}^{C} \prod_{u \in \mathcal{V}} q(s_{cu}, \boldsymbol{\alpha}^{(c,u)}) \prod_{i:D_i=0} q(Y_i) \prod_{i=1}^{N} q(Z_i) \prod_{c=1}^{C} \prod_{\ell=1}^{L} q(\rho_{c\ell}) \prod_{u \in \mathcal{V}^*} \prod_{j,k} q(\gamma_{jk}^{(u)}). \qquad (4.16)$$

By the well-known equality, $\log \mathrm{pr}(\mathcal{D}) = \mathcal{E}(q) + KL(q||\mathrm{pr}(\boldsymbol{\Gamma} \mid \mathcal{D}))$. Because $\log \mathrm{pr}(\mathcal{D})$ is constant in q, minimizing the KL divergence between the variational family and the true posterior distribution is equivalent to maximizing $\mathcal{E}(q)$, or "evidence lower bound (ELBO)" which is defined by $\mathcal{E}(q) = \int q(\boldsymbol{\Gamma}) \log \frac{\mathrm{pr}(\mathcal{D},\boldsymbol{\Gamma})}{q(\boldsymbol{\Gamma})} d\boldsymbol{\Gamma}$ where $\boldsymbol{\Gamma}$ collects all the unknowns. We further bound $\mathcal{E}(q)$ from below by bounding terms in $\mathrm{pr}(\mathcal{D}, \boldsymbol{\Gamma})$ that involve sigmoid functions that create non-conjugacy issues under the Gaussian-distributed priors used in this paper hindering simple closed-form VI updates. In particular, following Jaakkola and Jordan (2000), we can bound

Equation (3.6) and $\sigma(X_{ij}^* \beta_{jk}^{(c)})$ from below respectively by

$$\lambda_k^{(c,g)} \geqslant \{h(\eta_k^{(c,g)}; \phi_k^{(c,g)})\}^{\mathbf{1}\{k<K\}} \prod_{s<k} h(-\eta_s^{(c,g)}; \phi_s^{(c,g)}), g = 0, 1, \ldots, G, \text{ and} \tag{4.17}$$

$$\sigma(X_{ij}^* \beta_{jk}^{(c)}) \geqslant h(X_{ij}^* \beta_{jk}^{(c)}; \psi_{jk}^{(c)}), \tag{4.18}$$

where we have used the inequality

$$h(x, \psi) := \sigma(\psi) \exp\{(x - \psi)/2 - g(\psi)(x^2 - \psi^2)\} \leqslant \sigma(x), \tag{4.19}$$

with $g(\psi) = \frac{1}{2\psi}[\sigma(\psi) - \frac{1}{2}]$ where ψ is a tuning parameter. As a result, the right-hand-side terms in Equations (4.17) and (4.18) are quadratic in $\alpha_k^{(c,u)}$ and $\gamma_{jk}^{(u)}$, paving the way for closed-form VI updates. Also see Durante, Rigon *and others* (2019) for a modern view of the technique as a bona fide variational algorithm with Pólya-Gamma augmentation. We now have a lower bound $\mathcal{E}^*(q)$ of $\mathcal{E}(q)$ which is defined as

$$\mathcal{E}^*(q) := \int q(\mathbf{\Gamma}) \log \frac{H(\mathcal{D}, \mathbf{\Gamma}; \boldsymbol{\psi}, \boldsymbol{\phi})}{q(\mathbf{\Gamma})} d\mathbf{\Gamma} \leqslant \int q(\mathbf{\Gamma}) \log \frac{\mathrm{pr}(\mathcal{D}, \mathbf{\Gamma})}{q(\mathbf{\Gamma})} d\mathbf{\Gamma} = \mathcal{E}(q), \tag{4.20}$$

where $H \leqslant \mathrm{pr}(\mathcal{D}, \mathbf{\Gamma})$ is obtained by applying the lower bounds in Equations (4.17) and (4.18) to relevant terms in (3.15) and has tuning parameters $\boldsymbol{\psi} := \{\psi_{jk}^{(c)}, j \in [J], k \in [K], c \in [C]\}$ and $\boldsymbol{\phi} := \{\phi_k^{(c,g)}, c \in [C], g \in \{0\} \cup [G], k \in [K-1]\}$; see Appendix B in the Supplementary Materials for the exact formula for calculation.

The variational algorithm then finds the optimal variational distribution in the variational family that maximizes $\mathcal{E}^*(q)$. In particular, we take the logarithm of the lower bound H of the joint probability density for data and unknowns with respect to a variational distribution q: $\mathbb{E}_q[\log H]$. The algorithm updates each factor in order while holding the rest fixed. The update for the j-th factor in the variational distribution is $\mathbb{E}_{q_{-j}}[\log H]$, where $\mathbb{E}_{q_{-j}}$ means taking expectations with respect to q over all but the variables in the j-th factor in q. The logarithmic of H can be written as in Appendix B in the Supplementary Materials. A desirable property of $\log H$ is that integration of $\log H$ with respect to each factor of q is in closed-form, which is a key ingredient of each VI update. The pseudo-code for the VI updates are provided in Algorithm 1. See Appendix A in the Supplementary Materials for the details of each update.

Cause-specific domain dissimilarity measure The framework motivates the following the estimated cophenetic distance (e.g., Sneath, Sokal *and others*, 1973) between the target domain and each of the source domains. In particular, the dissimilarity measure is $dist_{\mathcal{T}_{w'}}(\mathsf{target} = 0, \mathsf{source} = g; \mathsf{cause} = c)$, $g = 1, \ldots, G$,

$c = 1, \ldots, C$, where $w'(P)$ is defined as the sum of all the modified edge weights along the path P connecting the target domain and source domain g in $\mathcal{T}_{w'}$, and the modified weight of an edge $(pa(u) \rightarrow u)$ is $w'_u = p_{cu} \cdot w_u$ for node $u \in \mathcal{V}$ in the domain tree (see Equation (A5) of VI updates in Appendix A in the Supplementary Materials for definition of p_{cu} (on a logit scale) as the variational approximation to $\mathbb{P}(s_{cu} = 1 \mid \mathcal{D})$).

Choice of K   We follow Bishop (2006) and use criterion $\mathcal{E}_K^*(q) + \log(K!)$ where $\mathcal{E}_K^*(q)$ is the lower bound of log marginal data likelihood for a K-class model and the correction term is to make different models comparable (e.g., Grimmer, 2011, Section 5.2).

Software   A free and publicly available `R` package that implements the VI algorithm for scalable approximate posterior inference is freely available at `https://github.com/zhenkewu/doubletree`. The package is designed to work under all possible patterns of observed and missing causes of death: (Scenario i) Y_i is missing in a single domain (say g): i-1) none has confirmed causes in domain g; i-2) there exists at least one observed Y_i in domain g; (Scenario ii) Y_i is missing in $M \geqslant 2$ domains; (Scenario iii) no missing Y_i in any domain. This paper has focused on Scenario (i-1) for simpler exposition.

5. SIMULATION STUDIES

We conduct two sets of simulation studies to evaluate the operating characteristics of the proposed method and demonstrate its better capability of estimating the target-domain CSMFs and assigning individual-level causes of death relative to a few alternatives with ad hoc specifications of information pooling across the domains. In the first set of simulations, we simulate data based on true parameters values under NLCM. In the second set of simulations, we use a validation data set and selectively mask a subset of deaths' true causes in a synthetically constructed target domain and then apply NLCM. The simulation designs, performance metrics, and results are detailed below.

Performance Metrics   First, we assess the overall accuracy by the so-called "CSMF accuracy" (Murray *and others*, 2011*b*), widely used as the metric to compare the estimated CSMF vector against the truth. The CSMF accuracy metric measures the L_1-distance between the estimated and true vectors of CSMFs, and is

normalized to range between 0 (worst) and 1 (best). It is defined as $\mathsf{CSMF}_{\mathsf{acc}}(\widehat{\pi}^{(0)}) = 1 - \frac{\sum_{c=1}^{C} |\widehat{\pi}_c^{(0)} - \pi_c^{(0)}|}{2(1 - \min_c \pi_c^{(0)})}$,

where $\pi_c^{(0)}$ is the true CSMF for cause c that we set in simulation design (or calculated by the empirical

distribution in the synthetic target domain data in Simulation II below). This formulation is also known as

the normalized absolute error in the quantification learning literature (González *and others*, 2017). Finally,

for the accuracy of COD classifications, we will use top cause accuracy: the fraction of deaths with the true

CODs in the top predicted causes.

5.1 *Simulation I*

Design We simulated $R = 200$ independent replicate data sets for different total sample sizes ($N = 1000, 4000$). To illustrate, we use a domain tree $\mathcal{T}_w$ shown in Figure 2(a) with equal edge weights and

true domain leaf groups; $\mathcal{T}_w$ has $p_{\mathsf{leaf}} = 6$ leaves and 3 domain leaf groups. The leaf "0" is set to be the target

domain leaf. For each N, we set each domain's sample size to be approximately N/p_{leaf} for $g = 0, 1, \ldots, G$

(with rounding where needed) to investigate balanced leaves and set the sample size in domain g to be

approximately $\frac{1}{5}N/p_{\mathsf{leaf}}$ or $\frac{4}{5}N/p_{\mathsf{leaf}}$ with equal chances for mimicking unbalanced observations across the

domain leaves. Within domain g, we further assign deaths into C causes by independently sampling from

categorical distributions with CSMFs $\boldsymbol{\pi}_g$, $g = 0, 1, \ldots, G$. We then simulated multivariate binary response

data for different dimensions $J = 20, 60$, for $K = 2$ classes according to Equations (2.2) and (2.3). We con-

sidered $C = 3$ causes. Two different sets of $\{\boldsymbol{\Theta}^{(c)}\}$ were considered; see Appendix D in the Supplementary

Materials for more details of the true parameter values and model setup.

For each simulated data set, we fitted the proposed model, based on which we compute the approximate

posterior mean of $\boldsymbol{\pi}^{(0)}$ obtained via its optimal variational distribution. In addition, we also compared against

a few NLCM-based approaches but with suboptimal, ad hoc specifications of information pooling between

the domains (to different degrees of cross-domain shrinkage). Figure 2(a) shows the domain groupings used

in these comparisons. In summary, we fit 1) the proposed method:"Domain Adaptive"; 2) "True Domain

Grouping": for any cause c, assume identical $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = g)$ for domains in a group (4 groups in

the simulation truth). To do so, for each c, we fix s_{cu} to 1s or 0s in a way that results in the true domain

grouping; 3) "Complete Pooling": completely ignore the external domain tree information during estimation

and also ignore the sample-to-domain mappings $\boldsymbol{D}$. By doing so, we assume $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = g)$ remains

the same between the domains; 4) "Ad hoc Domain Grouping": use a manual domain grouping that is finer than the true domain grouping; 5) "No Domain Grouping": same as 3) except D is used during estimation so that data are recognized to have come from different domains.

Results Figure 2(b) compares the methods in terms of CSMF accuracy. The proposed NLCM Domain Adaptive method adaptively learns the the domain groupings and produced the CSMF estimate for the target domain with the highest accuracy. The accuracy is comparable to the ones obtained under the true domain grouping. The method with complete pooling between domains generally performs the worst; this should not be surprising given the simulation truth is to mimic situations where $\mathrm{pr}(X_i \mid Y_i, D_i = g)$ differs by domain. NLCMs with ad hoc domain grouping and no grouping produced more accurate estimates than the one with complete pooling between the domains. However, because both methods are based on domain groupings that are finer than the true domain grouping, they do not fully use similar domains to improve the accuracy of estimating the conditional distributions of the VA responses given a cause, resulting in sizable losses in CSMF accuracy. Similar relative patterns are also clear when RMSEs are compared (see Figure Appendix Figure 3). In addition, as expected the conditional dependence modeled by the NLCMs for each cause improved the individual-level classification performance (top cause accuracy) relative to methods that ignored conditional dependence (results not shown here).

5.2   *PHMRC VA Data: Background and Description*

In Section Appendix D.1 and Section 6, we will use the Population Health Metrics Research Consortium (PHMRC) VA validation data with physician-coded causes of death to evaluate the performance of the proposed method. We first review aspects of the data set pertinent to our study.

PHMRC VA validation data collection was implemented in six sites in four countries: Andhra Pradesh ("AP"), India; Uttar Pradesh ("UP"), India; Dar es Salaam ("Dar"), Tanzania; Pemba Island ("Pemba"), Tanzania; Bohol, Philippines; Mexico City ("Mexico"), Mexico. The goal is to create a high quality validation data set from different populations to evaluate comparative method performance and make recommendations for future VA implementations. See Murray *and others* (2011*a*) for a more complete description of the PHMRC VA validation data. The data set with 7,841 adult deaths and 168 symptoms that we use here is further based on preprocessing in McCormick *and others* (2016).

Figure 3 shows the domain hierarchy via a rooted tree with six leaves at the top margin. The domain hierarchy uses country membership information to form the leaves and internal nodes. Unless otherwise stated, we set all edge weights to be one for illustration. External domain-level information that is highly associated with CSMFs may also be incorporated to form the domain hierarchy, e.g., via hierarchical clustering of a variety of domain-level information that may alter symptom-cause relationships, such as time periods, level of VA interviewer training, and differential availability of treatments that mitigate a subset of symptoms interviewed in VA. As a secondary feature of the proposed method, the left margin of Figure 3 shows the cause hierarchy with 35 leaves along with coarser aggregated cause definitions.

Among the 168 symptoms, 63 have a missing rate of higher than 1%, of which 37 has a missing rate higher than 5%. The highest missing rate is 96.6% for: "Was there pain in the upper belly?"; the next highest missing rate is 92%, for four questions related to where rash was located if present: "Trunk?","Extremities?","Everywhere?", "Other locations?" In addition, the numbers of missing symptoms for a subject are between 1 and 76, with a median of 22. Missing data in VA in the form of "Don't Know" or "Choose not to answer" appear for various reasons. Missing data have been considered by Kunihama *and others* (2020) under the assumption of missing at random. Absent additional information regarding individual-specific missing data mechanism and given the lack of likely alternative sensitivity assumptions about missing data, we will assume missing at random in this paper when conducting model estimation and method comparisons. This sets the stage for fair relative comparisons with other existing methods that either assume missing completely at random or missing at random. In our proposed model, because given a cause and a class, symptoms are assumed mutually independent, when calculating the cause-specific likelihood for an individual in class k with a subset of missing symptoms, we simply do products of Bernoulli likelihoods only over symptoms with non-missing information; see the implementation in Steps 1a and 1e in Appendix A during variational updates.

5.3   *Simulation II: Semi-Synthetic*

Here we use the PHMRC VA validation data to evaluate the performance of the proposed method. Because validation data contains gold-standard causes of death, we split the original data set into source and target domain data sets and mask the causes of the deaths allocated to the target domain. Because we use the real PHMRC data while masking a subset of causes of death, we refer to this simulation as "semi-synthetic".

See Appendix D.1 in the Supplementary Materials for the details about the design and results showing that domain adaptive estimation provides more accurate CSMF estimates and top cause COD assignment.

6. Domain Adaptation across Actual PHMRC Sites

This scenario uses the actual PHMRC domain designation: one site as the target domain with the gold-standard cause-of-death labels masked, the rest five sites as source domains with observed gold-standard causes-of-death. Table 1 shows the CMSF accuracy when each of the six PHMRC study sites is treated as the target domain iteratively. The method selected two-class models. The domain adaptive approach achieved better accuracies in CSMF estimation and slight improvements on the top-cause classification accuracies. The somewhat low accuracies using PHMRC data are well known, motivating new ongoing validation data collection by our substantive collaborators based on which we will further test our method. To illustrate how to interpret results from the domain adaptive method, we pick AP as the target domain. In addition, we pick six causes (out of 35 causes used during model fittings) to illustrate the results of cause-specific shrinkage. The causes are selected based on the different levels of shrinkage: near-complete pooling (cause "Drowning"), and no substantial shrinkage between the domains (cause AIDS, Stroke, Renal Failure, Tuberculosis (TB), Inflammatory Heart Disease (IHD)-Acute Myocardial Infarction (MI)).

Using AP as an example target domain, Figure 4 shows two-class LCM results for some pairs of (cause,domain). For each of the six causes shown in Figure 4(b), the relative importance of the two classes varies greatly by domain, indicating substantive differences in $\text{pr}(\boldsymbol{X}_i \mid Y_i = c, D_i = g)$ between the domains. This is evident from the patterns of the class mixing weights for "TB" and "Renal Failure". Interestingly, this is not uniformly true for all causes, mostly notably for "Drowning", which placed large weights on the first class regardless of domain. A plausible explanation for this empirical finding is that symptoms are easily recognizable and highly specific if a death is caused by drowning. This analysis also empirically confirms that the conditional distributions of VA responses given each cause may vary by cause.

In addition, we also calculate the estimated cophenetic distance (e.g., Sneath, Sokal *and others*, 1973) between the target domain and each of the source domains (see Section 4). Figure 4(c) shows for each of 35 causes (rows), five estimated distances with smaller values representing higher similarity between the target domain (AP) and each of the five source domains shown on the x-axis. We make a few interesting empirical

observations. First, UP is estimated to be most similar to the target domain for almost all causes, which is perhaps unsurprising given AP and UP are two states in India. Second, the joint pattern of similarity between AP and five source domains differ by cause. For example, for causes like "Drowning", "Malaria", "Asthma", the dissimilar measures are all estimated to be small, indicating AP is uniformly similar to all other source domains in terms of symptom-cause conditional distributions. This uniformity in the dissimilarity measure may be explained by reporting of specific symptoms that vary little by domain. On the other hand, for causes with complex etiologies and manifest symptoms such as "Renal Failure", the target domain AP is estimated to be most similar to UP and least so to Bohol.

7. Discussion

Summary   In this paper, we presented a hierarchical Bayesian approach to use individual-level multivariate binary responses obtained from a target domain in the absence of any gold-standard categorical labels ("causes" in VA) for the estimation of the target population fractions of the label categories ("CSMFs" in VA). This is made possible by using individual-level data from multiple source domains where additional gold-standard cause labels are available. The data from multiple domains are integrated following a data-driven tree-structured shrinkage approach, so that for each cause, domains that have similar conditional distributions of the responses given the cause are encouraged to be pooled to improve estimation. We achieved this goal via a logistic stick-breaking Gaussian diffusion process on the mixing weights along a pre-specified domain hierarchy. In addition, an analyst may use another hierarchy applied to causes to regularize the parameter estimation when characterizing the conditional distributions of the responses given a cause. Simulation studies show that the proposed method produces more accurate estimation of CSMFs than methods that either ignore between-domain differences by complete pooling of data across the domains, ad hoc specification of domain grouping, or no domain grouping at all; the proposed method performs similarly to an oracle method with the true domain groupings. Although this paper focuses on the more challenging case where no cause label is observed in the target domain for simpler exposition, the model readily generalizes to using gold-standard cause labels for a subset of deaths in the target domain. The software accompanying this paper has implemented such extensions.

Limitations Practical limitations of the proposed approach may exist. First, the hierarchy for the domain is pre-specified based on external domain-level information (geo-locations of the study sites in this paper) and are not estimated from the VA data themselves. Methods that specify a prior over the space of domain hierarchies for posterior inference may be fruitful at extra computational costs (e.g., Knowles and Ghahramani, 2014). Second, deviation from missing-at-random assumption for the VA questionnaire responses may impact the model results and performance during domain adaptations. This issue remains challenging and to be explored in collaboration with VA substantive experts to identify common and major reasons leading to missing data. Third, the cause list used in this paper is chosen based on clear clinical meanings of distinct etiologic implications and also for methodological illustration. Certain causes of death, such as COVID-19 related deaths, may quickly emerge as prominent causes in some populations that necessitates updated cause lists. Fourth, PHMRC data is currently the only validation data set for evaluation; future VA gold standard data sets will be available for additional validation. Fifth, additional unstructured narrative texts are available, and may improve the capacity of the present approach by augmenting the VA symptoms with absence or presence of derived text features (e.g., cause-discriminative words). Finally, the current results are agnostic to additional information about symptom-cause relationships. Estimation accuracy may be further improved by incorporating these information (e.g., McCormick *and others*, 2016; Schifeling, Reiter *and others*, 2016).

Future directions There are a few statistical extensions that may further improve the utility of the proposed method. First, additional individual-level covariates, such as age, pregnancy status, and seasonality, that may explain variation in the conditional distribution of the VA questionnaire responses given a cause $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i, D_i)$. Our framework readily incorporates discrete individual-level covariates via concatenation with the vector of VA questionnaire responses. For continuous individual-level covariates, Moran *and others* (2021) illustrated an approach based on a different framework of factor models, which however is not suited to dealing with a domain absent any cause-of-death label. Mixed outcome extensions are desirable (e.g., Zhang *and others*, 2021). Second, extensions that incorporate priors over K that may differ by cause can lead to posterior inference of cause-specific values of K. To this end, for each cause, sparsity priors over a probability simplex may be introduced to encourage absence of a subset of classes in certain domains. Third, latent class model for $\mathrm{pr}(\boldsymbol{X}_i \mid Y_i, D_i)$ is a simple example of probabilistic tensor decomposition for multivariate discrete data, which can be replaced with alternatives against which comparisons are warranted (e.g., Bhattacharya

and Dunson, 2012; Zhou *and others*, 2015; Gu *and others*, 2021). Fourth, negative transfer issues have been noted in machine learning literature on transfer learning (e.g., Pan and Yang, 2009; Zhang *and others*, 2020). Our approach is based on an assumption of a shared set of class-specific response profiles $\Theta^{(c)}$ for each cause to facilitate interpretation. It is of interest to evaluate potential impact of deviations from such an assumption and to study mitigating solutions (e.g., Stephenson, Herring and Olshan, 2020). We leave these topics for future research.

Supplementary Material

Supplementary material is available online.

Acknowledgments

References

Bhattacharya, A. and Dunson, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association* **107**(497), 362–377.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**(518), 859–877.

Carbonetto, P., Stephens, M. *and others*. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* **7**(1), 73–108.

Chandramohan, D., Fottrell, E., Leitao, J., Nichols, E., Clark, S. J., Alsokhn, C., Munoz, D. C., AbouZahr, C., Pasquale, A. D., Mswia, R., Choi, E., Baiden, F., Thomas, J., Lyatuu, I., Li, Z. R., Larbi-Debrah, P., Chu, Y., Cheburet, S., Sankoh, O., Bad, A. M., Fat, D. M., Setel, P., Jakob, R. *and others*. (2021). Estimating causes of death where there is no medical certification: Evolution and state of the art of verbal autopsy. *In Press, Global Health Action*.

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.

DUNSON, D. B. AND XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104**(487), 1042–1051.

DURANTE, D., RIGON, T. *and others*. (2019). Conditionally conjugate mean-field variational Bayes for logistic models. *Statistical Science* **34**(3), 472–485.

FIKSEL, J., DATTA, A., AMOUZOU, A. AND ZEGER, S. (2021). Generalized Bayes quantification learning under dataset shift. *Journal of the American Statistical Association*, 1–19.

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.

GRIMMER, J. (2011). An introduction to Bayesian inference via variational approximations. *Political Analysis* **19**(1), 32–47.

GU, Y., EROSHEVA, E. A., XU, G. AND DUNSON, D. B. (2021). Dimension-grouped mixed membership models for multivariate categorical data. *arXiv preprint arXiv:2109.11705*.

JAAKKOLA, T. S. AND JORDAN, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**(1), 25–37.

KING, G. AND LU, Y. (2008). Verbal autopsy methods with multiple causes of death. *Statistical science* **23**(1), 78–91.

KNOWLES, D. A. AND GHAHRAMANI, Z. (2014). Pitman yor diffusion trees for bayesian hierarchical clustering. *IEEE transactions on pattern analysis and machine intelligence* **37**(2), 271–289.

KUNIHAMA, T., LI, Z. R., CLARK, S. J., MCCORMICK, T. H. *and others*. (2020). Bayesian factor models for probabilistic cause of death assessment with verbal autopsies. *Annals of Applied Statistics* **14**(1), 241–256.

LAZARSFELD, P. F. (1950). *The logical and mathematical foundations of latent structure analysis*, Volume IV, Chapter The American Soldier: Studies in Social Psychology in World War II. Princeton, NJ: Princeton University Press, pp. 362–412.

LI, M., PARK, D. E., AZIZ, M., LIU, C. M., PRICE, L. B. AND WU, Z. (2021a). Integrating sample similarities into latent class analysis: A tree-structured shrinkage approach. *Biometrics*, In press.

LI, Z. R., MCCOMICK, 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.

LI, Z. R., THOMAS, J., CHOI, E., MCCORMICK, T. H. AND CLARK, S. J. (2021b). The openVA toolkit for verbal autopsies. *arXiv preprint arXiv:2109.08244*.

LI, Z. R., WU, Z., CHEN, I. AND CLARK, S. J. (2021c). Bayesian nested latent class models for cause-of-death assignment using verbal autopsies across multiple domains. *arXiv preprint*.

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.

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)*.

MURRAY, C. J., LOPEZ, A. D., BLACK, R., AHUJA, R., ALI, S. M., BAQUI, A., DANDONA, L., DANTZER, E., DAS, V., DHINGRA, U. *and others*. (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. *Popul Health Metr* **9**(1), 28.

ORMEROD, J. T. AND WAND, M. P. (2010). Explaining variational approximations. *The American Statistician* **64**(2), 140–153.

PAN, S. J. AND YANG, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**(10), 1345–1359.

SCHIFELING, T. A., REITER, J. P. *and others*. (2016). Incorporating marginal prior information in latent class models. *Bayesian Analysis* **11**(2), 499–518.

SNEATH, P. H., SOKAL, R. R. *and others*. (1973). *Numerical taxonomy. The principles and practice of numerical classification.*.

STEPHENSON, B. J. K., HERRING, A. H. AND OLSHAN, A. (2020). Robust clustering with subpopulation-specific deviations. *Journal of the American Statistical Association* **115**(530), 521–537.

THOMAS, E. G., TRIPPA, L., PARMIGIANI, G. AND DOMINICI, F. (2020). Estimating the effects of fine particulate matter on 432 cardiovascular diseases using multi-outcome regression with tree-structured shrinkage. *Journal of the American Statistical Association* **115**(532), 1689–1699.

TITSIAS, M. AND LÁZARO-GREDILLA, M. (2011). Spike-and-slab variational inference for multi-task and multiple kernel learning. *Advances in Neural Information Processing Systems* **24**, 2339–2347.

TÜCHLER, R. (2008). Bayesian variable selection for logistic models using auxiliary mixture sampling. *Journal of Computational and Graphical Statistics* **17**(1), 76–94.

WORLD HEALTH ORGANIZATION. (2021). Who civil registration and vital statistics strategic implementation plan 2021-2025.

ZHANG, W., DENG, L., ZHANG, L. AND WU, D. (2020). Overcoming negative transfer: A survey. *arXiv preprint arXiv:2009.00909*.

ZHANG, Z., NISHIMURA, A., BASTIDE, P., JI, X., PAYNE, R. P., GOULDER, P., LEMEY, P. AND SUCHARD, M. A. (2021). Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models. *The Annals of Applied Statistics* **15**(1), 230–251.

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.
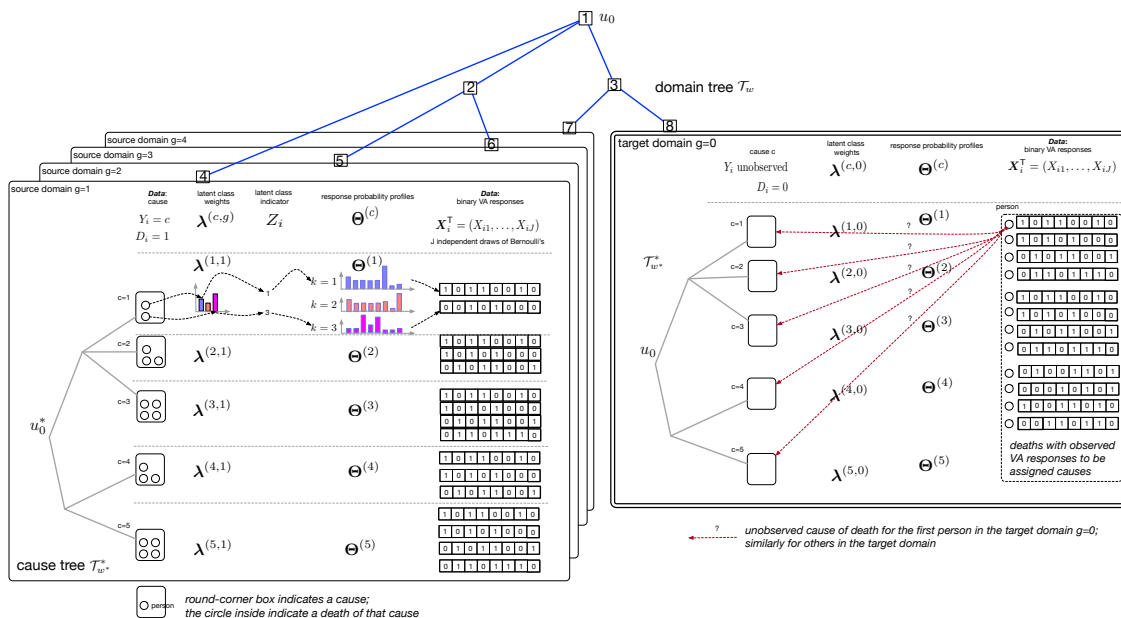
Fig. 1: Schematic representation of the nested LCM model structure. *Top*) An eight-node (root $u = u_0$) tree over five hypothetical domains is used to specify a tree-structured shrinkage prior for $\boldsymbol{\lambda}^{(c,g)}, g = 0, 1, \ldots, G$; *Left*) $G = 4$ source domains ($D_i = 1, ..., 4$), shown in overlaid plates; causes of deaths are observed to be in one of $C = 5$ hypothetical causes. Hypothetical observed $J = 8$ binary VA responses are also shown; *Right*) one target domain where the causes of deaths Y_i's are unobserved but the binary VA responses $\boldsymbol{X}_i$'s are observed. The cause hierarchy, as a secondary feature, is represented by a tree with five leaves representing $C = 5$ causes. Three latent classes ($K = 3$) are illustrated here.

Table 1: CSMF accuracy when each of the six PHMRC study sites is treated as the target domain iteratively.

Method	Bohol	Mexico	AP	UP	Dar	Pemba
			CSMF Accuracy			
Domain Adaptive	0.68	0.72	0.70	0.66	0.66	0.63
Complete Pooling	0.67	0.67	0.67	0.61	0.61	0.58
Ad Hoc Domain Grouping	0.67	0.68	0.68	0.61	0.65	0.62
No Domain Grouping	0.67	0.64	0.66	0.62	0.63	0.60
			Top Cause Accuracy			
Domain Adaptive	0.32	0.30	0.34	0.36	0.34	0.41
Complete Pooling	0.32	0.29	0.35	0.36	0.33	0.39
Ad Hoc Domain Grouping	0.32	0.30	0.36	0.35	0.33	0.38
No Domain Grouping	0.32	0.30	0.34	0.35	0.33	0.40
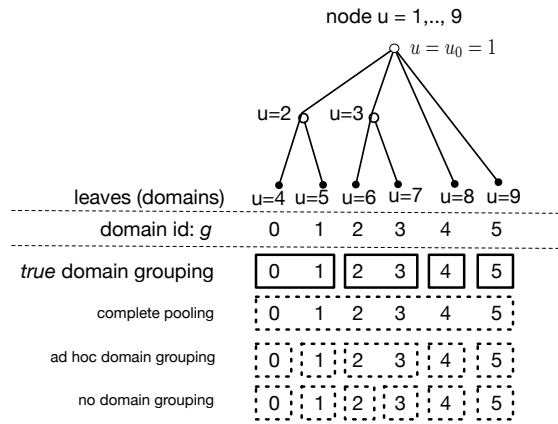
(a) Simulation I: domain tree and different domain groupings used in comparison.
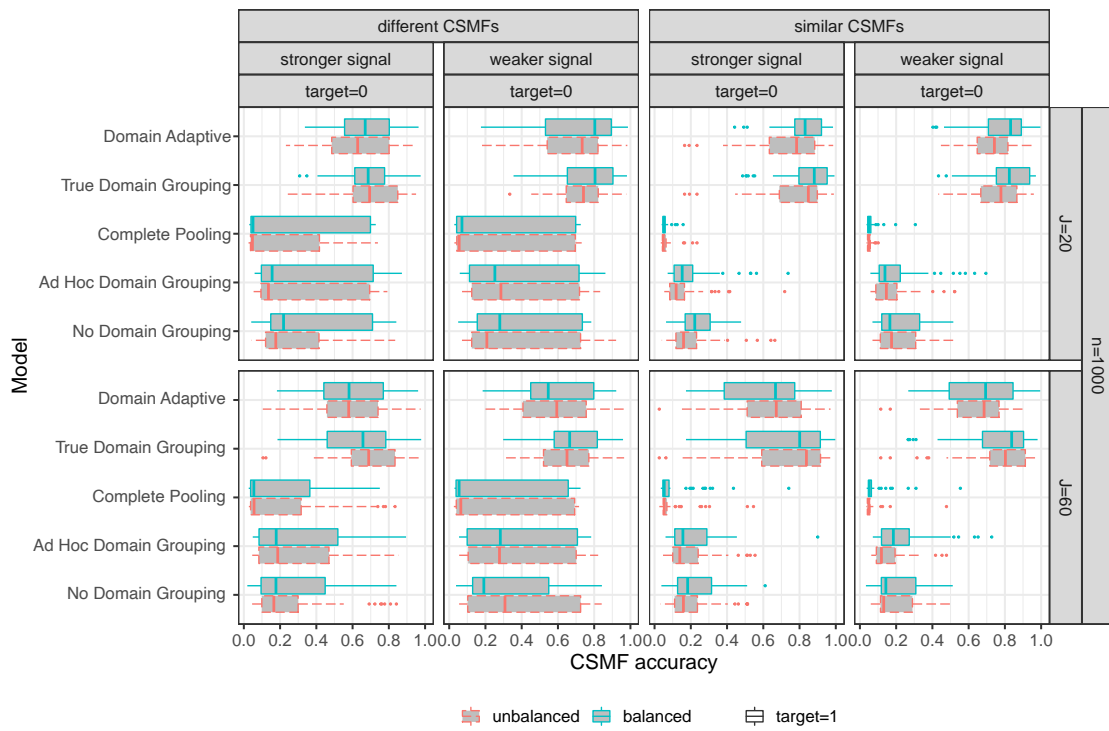
(b) Simulation I: CSMF accuracy comparison.

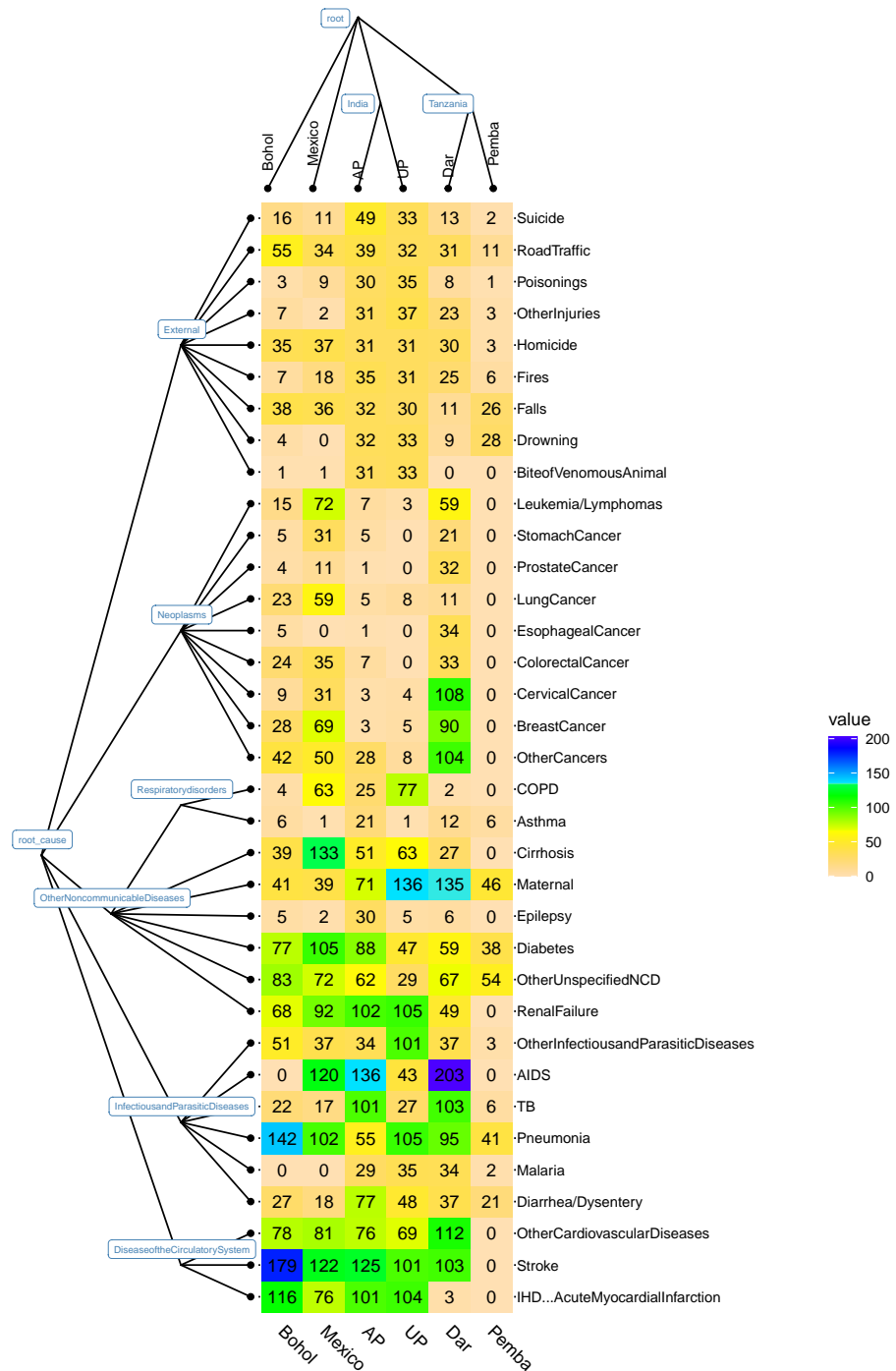Fig. 2: Simulation I: domain tree setup and results.

Fig. 3: Death counts by cause and site for $N = 7,841$ deaths and $J = 168$ across all six sites in the PHMRC data set. The exact death counts are shown in corresponding cells. Shown on the left and top margins are the cause and domain hierarchies assumed in the data analysis. We will mask the causes-of-death in one site during method testing so the site with masked causes is the target domain and the rest sites are source domains. The domains closer in the domain hierarchy (top margin) are *a priori* more likely fused to have the same vector of class mixing weights. The proposed method has a secondary feature that can incorporate a cause hierarchy (left margin) with 35 causes on the leaves and six aggregated causes represented by internal nodes.
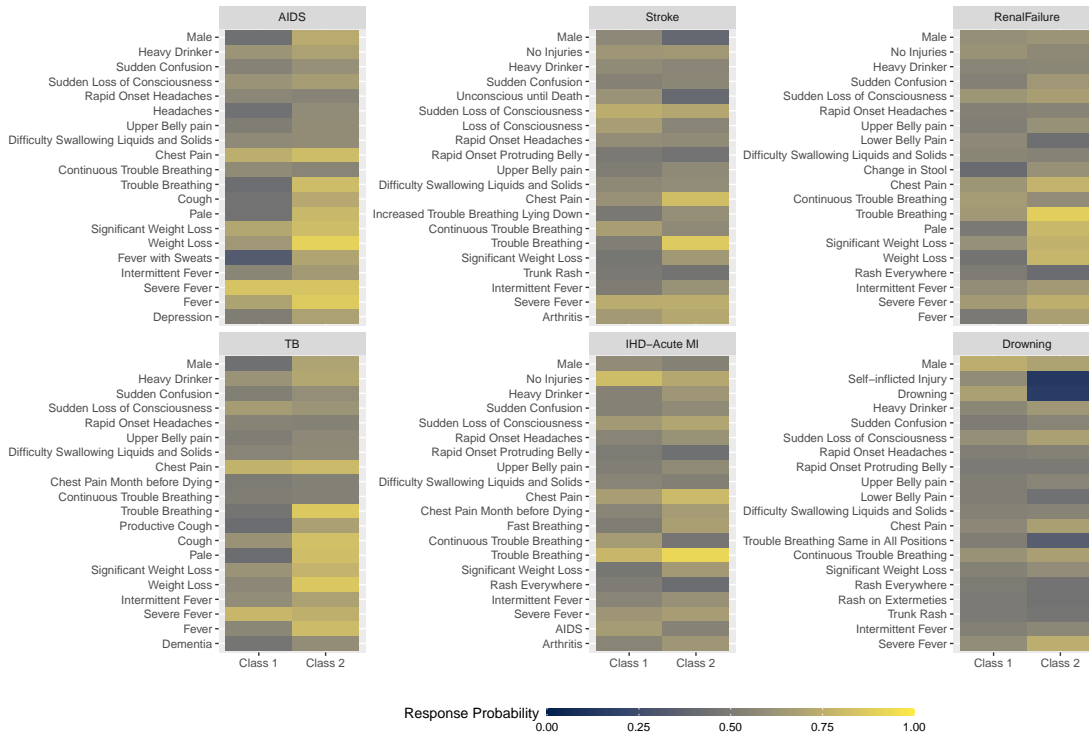
(a) Class-specific response probabilities based on a $K = 2$ class model (top 5 causes in AP and Drowning; top 20 symptoms with the highest estimated marginal probabilities).
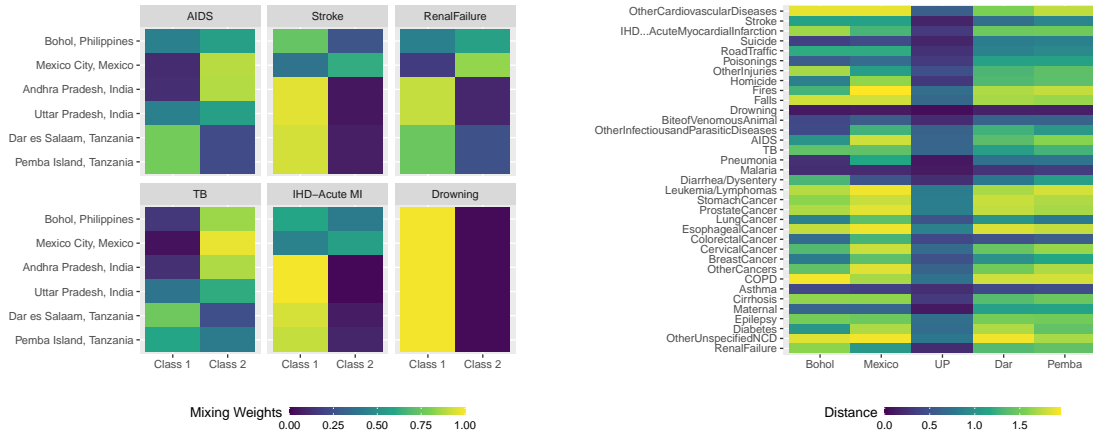
(b) Variation of class-mixing weights between domains; six sets of weights are shown for six causes of deaths (the model uses 35 causes).

(c) Estimated cause-specific cophenetic distances between AP (target) and each of the five source domains; 35 rows representing 35 causes used during model fitting.

Fig. 4: PHMRC results based on a $K = 2$ class nested LCM; for illustration, model results using AP as the target domain are shown here.

Algorithm 1: Pseudocode of Variational Bayes Algorithm

Data:

   (a) Multivariate binary data $\mathbf{X}$

   (c) The domain ids $\boldsymbol{D}$;

   (b) The cause ids Y_i for subject i with $D_i \neq 0$;

   (d) A weighted rooted tree for domains $\mathcal{T}_w = (\mathcal{T} = (\mathcal{V}, E), w)$: leaves $\mathcal{G} = \{0\} \cup [G] \subset \mathcal{V}$, edge lengths $\boldsymbol{w} = (w_u)_{u \in \mathcal{V}}$;

   (e) A weighted rooted tree for causes $\mathcal{T}_w^* = (\mathcal{T}^* = (\mathcal{V}^*, E^*), w^*)$: leaves $[C] \subset \mathcal{V}^*$;

Fixed Hyperparameters:

   (a′) The number of classes $K \geqslant 2$; levels $\ell_u \in [L]$ for all nodes $u \in \mathcal{V}$; levels $\ell_u^* \in [L^*]$ for all nodes $u \in \mathcal{V}^*$;

   (b′) Hyperparameters for the prior probability of $s_{cu} = 1$: $(a_{c\ell}, b_{c\ell})$, $c \in [C]$, $\ell \in [L]$;

Initialize:

   (a″) $t \longleftarrow 0$; Initialize $q_t(\boldsymbol{s}, \boldsymbol{\alpha})$ and $q_t(\boldsymbol{\gamma})$ `// (see Step 0 in Appendix A1)`

   (b″) Set an initial ELBO $\mathcal{E}_0^* \longleftarrow 0$

1 $t \longleftarrow 1$; $\mathcal{E}_1^* \longleftarrow \mathcal{E}_0^* + 2\epsilon$

2   **while** $|\mathcal{E}_t^* - \mathcal{E}_{t-1}^*| > \epsilon$ **do**

3 $q_t(\boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \longleftarrow q_{t-1}(\boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$

4 $\boldsymbol{\phi}^{(t)} \longleftarrow \boldsymbol{\phi}^{(t-1)}$; $\boldsymbol{\psi}^{(t)} \longleftarrow \boldsymbol{\psi}^{(t-1)}$

5 $\boldsymbol{\tau}_1^{(t)} \longleftarrow \boldsymbol{\tau}_1^{(t-1)}$; $\boldsymbol{\tau}_2^{(t)} \longleftarrow \boldsymbol{\tau}_2^{(t-1)}$

6   **for** $c \in \mathcal{C}$ **do**

7   **for** $i \in [N]$ **do**

8   **for** $k \in [K]$ **do**

9   **if** ($D_i == 0$) $e_{ic}^{(t)} \longleftarrow \mathrm{argmax}_{e_{ic}} \mathcal{E}_t^*(q)$ `// (See Step 1a in Appendix A1)`

10   **if** ($D_i \in \{1, \ldots, G\}$)

11 $r_{ik}^{(t)} \longleftarrow \mathrm{argmax}_{r_{ik}} \mathcal{E}_t^*(q)$ `// (See Step 1b in Appendix A1)`

12   **for** $g = 0, 1, \ldots, G$ **do**

13 $q_t(\boldsymbol{\pi}^{(g)}) \longleftarrow \mathrm{argmax}_{q_t(\boldsymbol{\pi}^{(g)})} \mathcal{E}^*(q)$ `// (See Step 1c in Appendix A1)`

14   **for** $c \in \mathcal{C}$ **do**

15   **for** $u \in \mathcal{V}$ **do**

16 $q_t(s_{cu}, \boldsymbol{\alpha}^{(c,u)}) \longleftarrow \mathrm{argmax}_{q_t(s_{cu}, \boldsymbol{\alpha}^{(c,u)})} \mathcal{E}_t^*(q)$ `// (see Step 1d in Appendix A1)`

17   **for** $u \in \mathcal{V}^*$ **do**

18 $q_t(\boldsymbol{\gamma}^{(u)}) \longleftarrow \mathrm{argmax}_{q_t(\boldsymbol{\gamma}^{(u)})} \mathcal{E}_t^*(q)$ `// (see Step 1e in Appendix A1)`

19   **for** $c \in \mathcal{C}$ **do**

20   **for** $\ell \in [L]$ **do**

21 $q_t(\rho_{c\ell}) \longleftarrow \mathrm{argmax}_{q_t(\rho_{c\ell})} \mathcal{E}_t^*(q)$ `// (see Step 1f in Appendix A1)`

22   **for** $k \in [K]$ **do**

                                                `// update local variational parameters for tighter lower bounds`

23   **for** $c \in \mathcal{C}$ **do**

24   **for** $g \in \mathcal{G}$ **do**

25 $\phi_k^{(c,g),(t)} \longleftarrow \mathrm{argmax}_{\phi_k^{(c,g)}} \mathcal{E}_t^*(q)$

26   **for** $j \in [J]$ **do**

27 $\psi_{jk}^{(c),(t)} \longleftarrow \mathrm{argmax}_{\psi_{jk}^{(c)}} \mathcal{E}_t^*(q)$ `// (see Step 2 in Appendix A1)`

28   **if** $t \bmod d = 0$ **then**

29   **for** $\ell \in [L]$ **do**

30 $\tau_\ell^{(t)} \longleftarrow \mathrm{argmax}_{\tau_l} \mathcal{E}_t^*(q)$

31   **for** $\ell \in [L^*]$ **do**

32 $\tau_\ell^{*,(t)} \longleftarrow \mathrm{argmax}_{\tau_\ell^*} \mathcal{E}_t^*(q)$ `// (see Step 3 in Appendix A1)`

33 $\mathcal{E}_t^* \longleftarrow ELBO(q_t)$ `// (see Step 4 in Appendix A1)`

34 $t \longleftarrow t + 1$

   Return: $q_{t-1}(\boldsymbol{s}, \boldsymbol{\alpha})$, $q_{t-1}(\boldsymbol{\gamma})$, $\{q_{t-1}(Y_i), D_i = 0\}$, $q_{t-1}(\boldsymbol{\varrho})$, $\{\mathcal{E}_1^*, \ldots, \mathcal{E}_{t-1}^*\}$