UNIVERSITY OF
MICHIGAN

Michigan Statistics for
Individualized Healthcare Lab
MiSIL

Tree-informed Bayesian Domain Adaptation

(“tree-guided between-domain information pooling”)

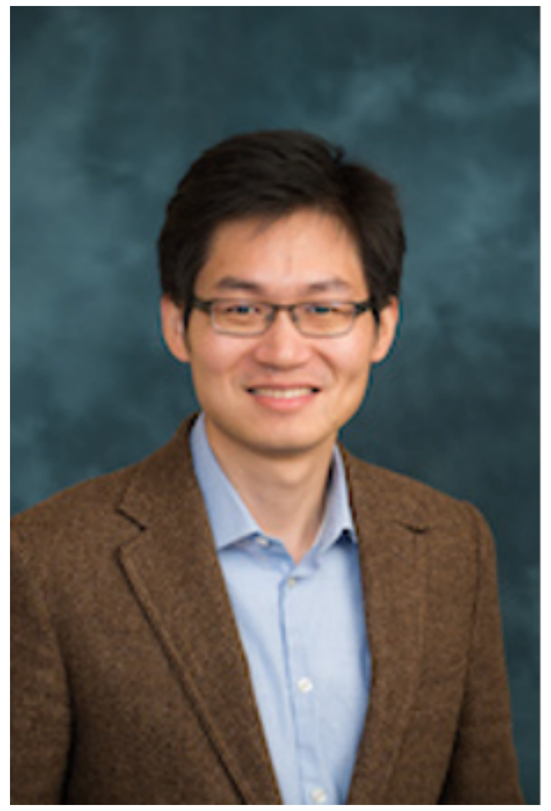
2022 ISBA Meeting
Montreal, Canada

Zhenke Wu, PhD

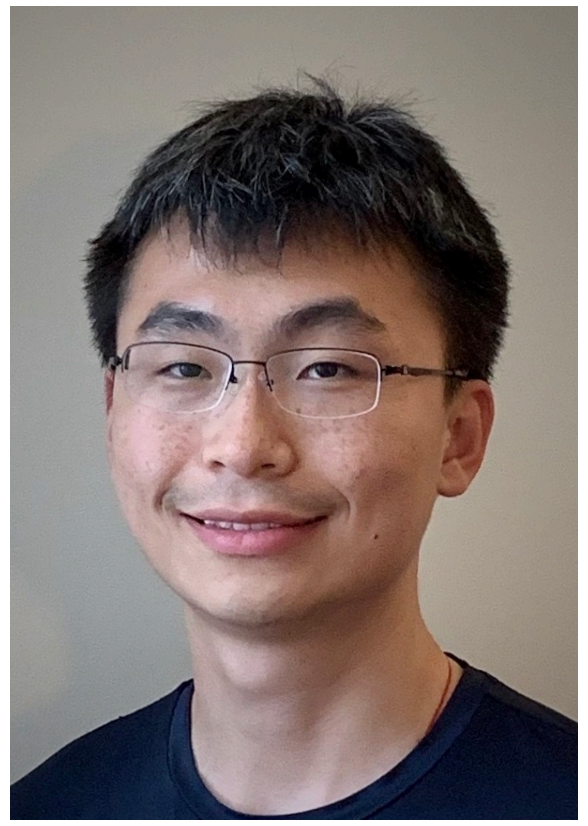
Assistant Professor of Biostatistics, University of Michigan
Research Assistant Professor of Michigan Institute for Data Science (MIDAS)
Michigan Statistics for Individualized healthcare Lab (MiSIL)

some slides sourced from Prof. Richard Li

Team

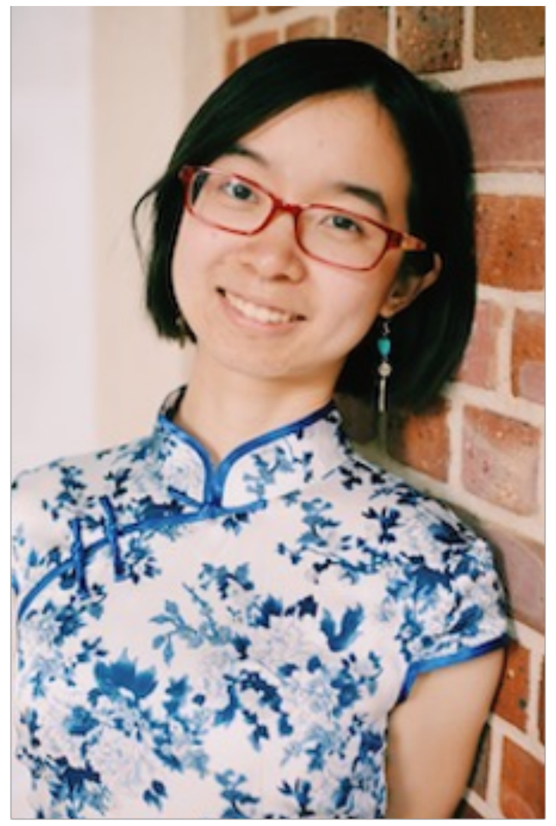


Zhenke Wu
Assistant Professor
of Biostatsitics, UMich



Zehang Richard Li
Assistant Professor
of Statistics,
UC Santa Cruz

PhD Candidate
Biostatistics, UMich

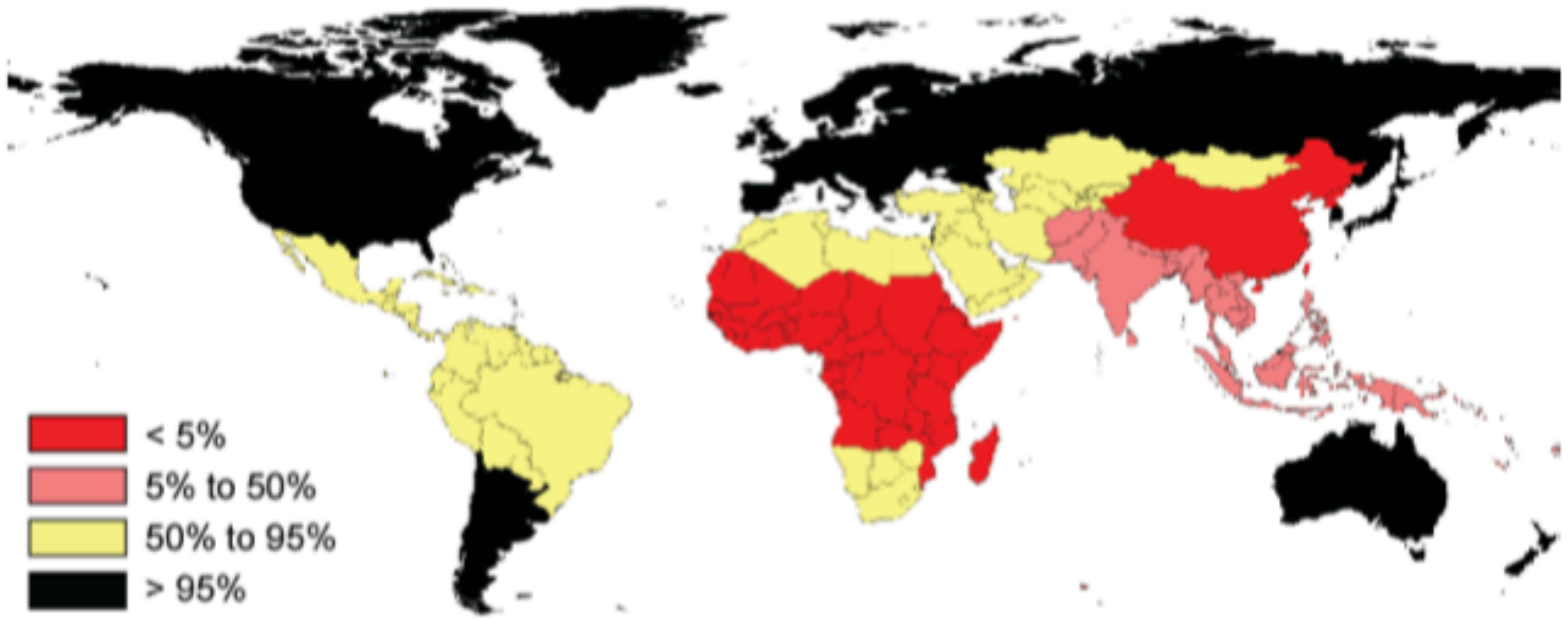


PhD Candidate
Biostatistics, UMich

Outline

- Part 1: Background
- Part 2: Proposed approach
    - Likelihood: nested latent class models
    - Prior for integrating tree info: Spike-and-slab logistic stick-breaking Gaussian diffusions along the tree (will spend less time on this today)
- Part 3: Results + software (🎄🎄)

“Hidden Deaths”



Proportions of deaths covered by vital registration.

Souce: Byass et al. (2013). *Reflections on the Global Burden of disease 2010 Estimates.* PLoS Med.

- Many people living in low- and middle-income countries are not covered by Civil Registration and Vital Statistics systems

- Cause-of-death data is lacking for 50% – 65% of the world’s population

- Registration of births and deaths, including cause of death information, is fundamental to any public health system.

Counting deaths

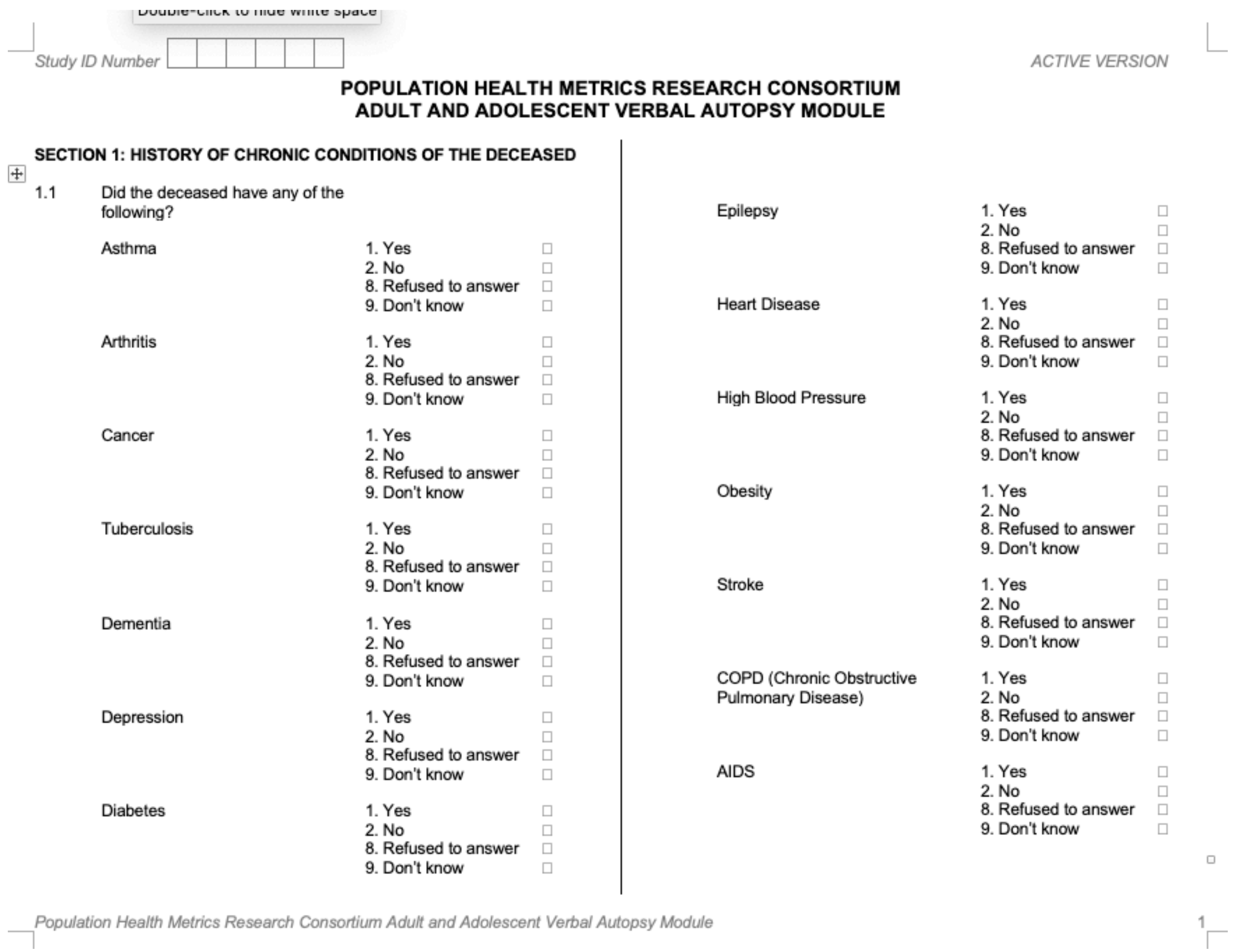
- Overall scientific goal:
  - Estimate cause-of-death distribution in the population and assign individual cause-of-death.
- Survey programs have been routinely used to obtain accurate demographic information such as births and deaths in low-resource settings
  - e.g., the Demographic and Health Surveys (DHS)
- Collecting information on cause-of-death (COD) is much harder.

Counting deaths: “The New Hope”

- Verbal autopsy (VA): interview relatives or caregivers and ask questions about the circumstances and symptoms leading up to a recent death.
- VA was first used in two research projects during 1965 – 1973 in Punjab, India.
- The use of VA has significantly expanded in the last five years.
- VA module has been integrated into the civil registration system in many countries.

Historical perspective and review: Chandramohan et al. (2022). Estimating causes of death where there is no medical certification: evolution and state of the art of verbal autopsy. *Global Health Actions.*

Population Health Metrics Research Consortium (PHMRC)
Verbal Autopsy Survey Form



Double-click to hide white space

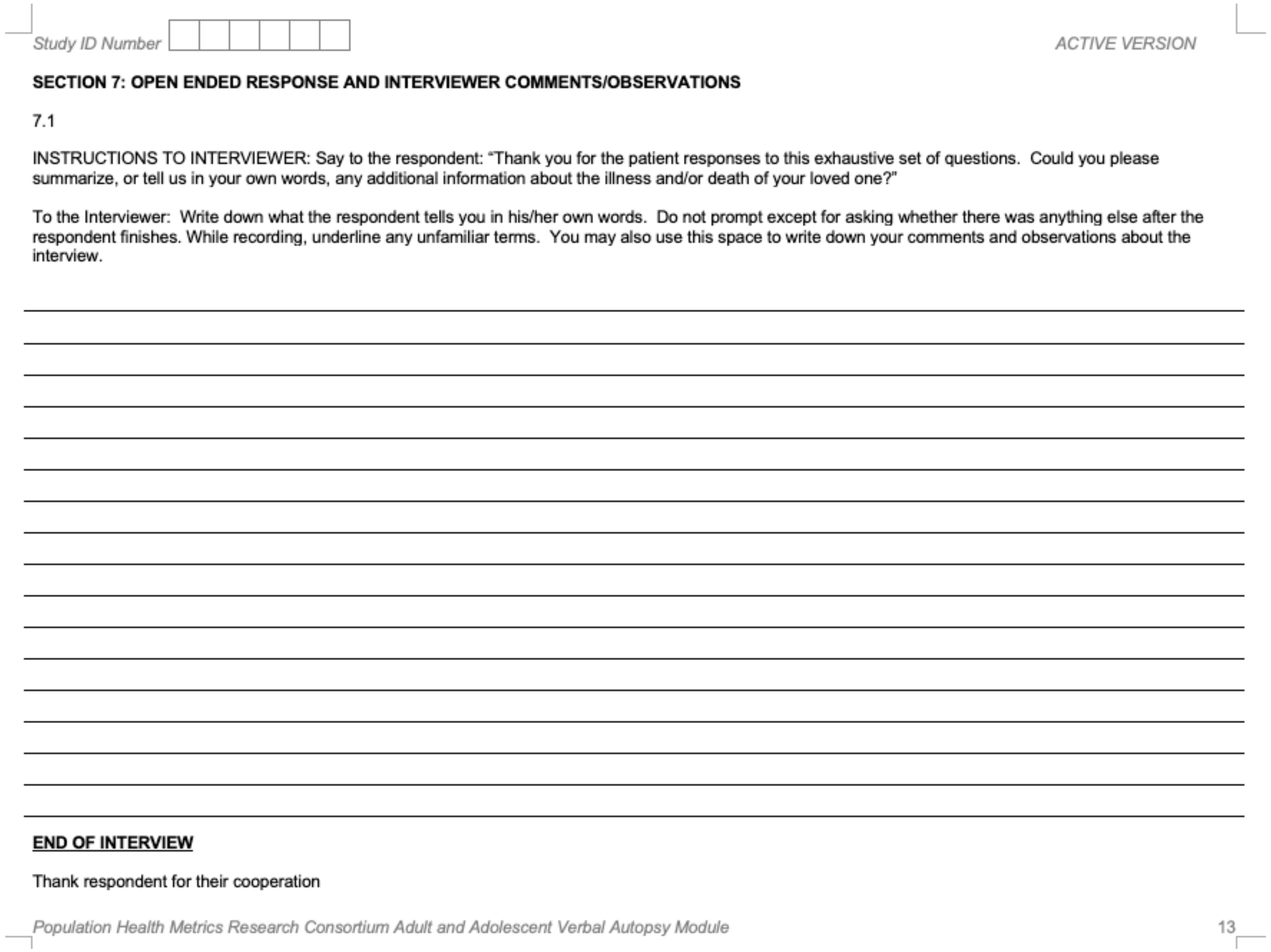
Study ID Number □□□□□□ ACTIVE VERSION

POPULATION HEALTH METRICS RESEARCH CONSORTIUM
ADULT AND ADOLESCENT VERBAL AUTOPSY MODULE

SECTION 1: HISTORY OF CHRONIC CONDITIONS OF THE DECEASED

1.1 Did the deceased have any of the following?

Asthma	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
Arthritis	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
Cancer	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
Tuberculosis	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
Dementia	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
Depression	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
Diabetes	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
Epilepsy	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
Heart Disease	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
High Blood Pressure	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
Obesity	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
Stroke	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
COPD (Chronic Obstructive Pulmonary Disease)	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □
AIDS	1. Yes □ 2. No □ 8. Refused to answer □ 9. Don't know □



Study ID Number □□□□□□ ACTIVE VERSION

SECTION 7: OPEN ENDED RESPONSE AND INTERVIEWER COMMENTS/OBSERVATIONS

7.1

INSTRUCTIONS TO INTERVIEWER: Say to the respondent: "Thank you for the patient responses to this exhaustive set of questions. Could you please summarize, or tell us in your own words, any additional information about the illness and/or death of your loved one?"

To the Interviewer: Write down what the respondent tells you in his/her own words. Do not prompt except for asking whether there was anything else after the respondent finishes. While recording, underline any unfamiliar terms. You may also use this space to write down your comments and observations about the interview.

END OF INTERVIEW

Thank respondent for their cooperation

2007-2010 Adult/Adolescent Module

typically 200-300 questions; some with complex skip patterns; implemented with varying qualities across sites; less costly and time-consuming than physician reviewing

Statistical Methods: “A Bayesian Revolution”

- Developments in analytic methods and reproducible open-source software for VA have greatly fostered confidence in large-scale implementations of VA in many low and middle income countries (LMICs).

  - Bayesian methods are critical: incorporate expert priors on symptom-cause relationships, uncertainty quantification

  - King and Lu (2008) Stat Sci.; McCormick et al. (2016) JASA; Li et al (2020) Bayesian Analysis; Moran et al. (2021) JRSS-C

  - openva.net (Clark, McCormick, Li and others): dedicated to open-sourcing stat tools for VA research

Li et al. (2021). The openVA toolkit for verbal autopsies. *arXiv:2109.08244*

“The Pain of Growth”

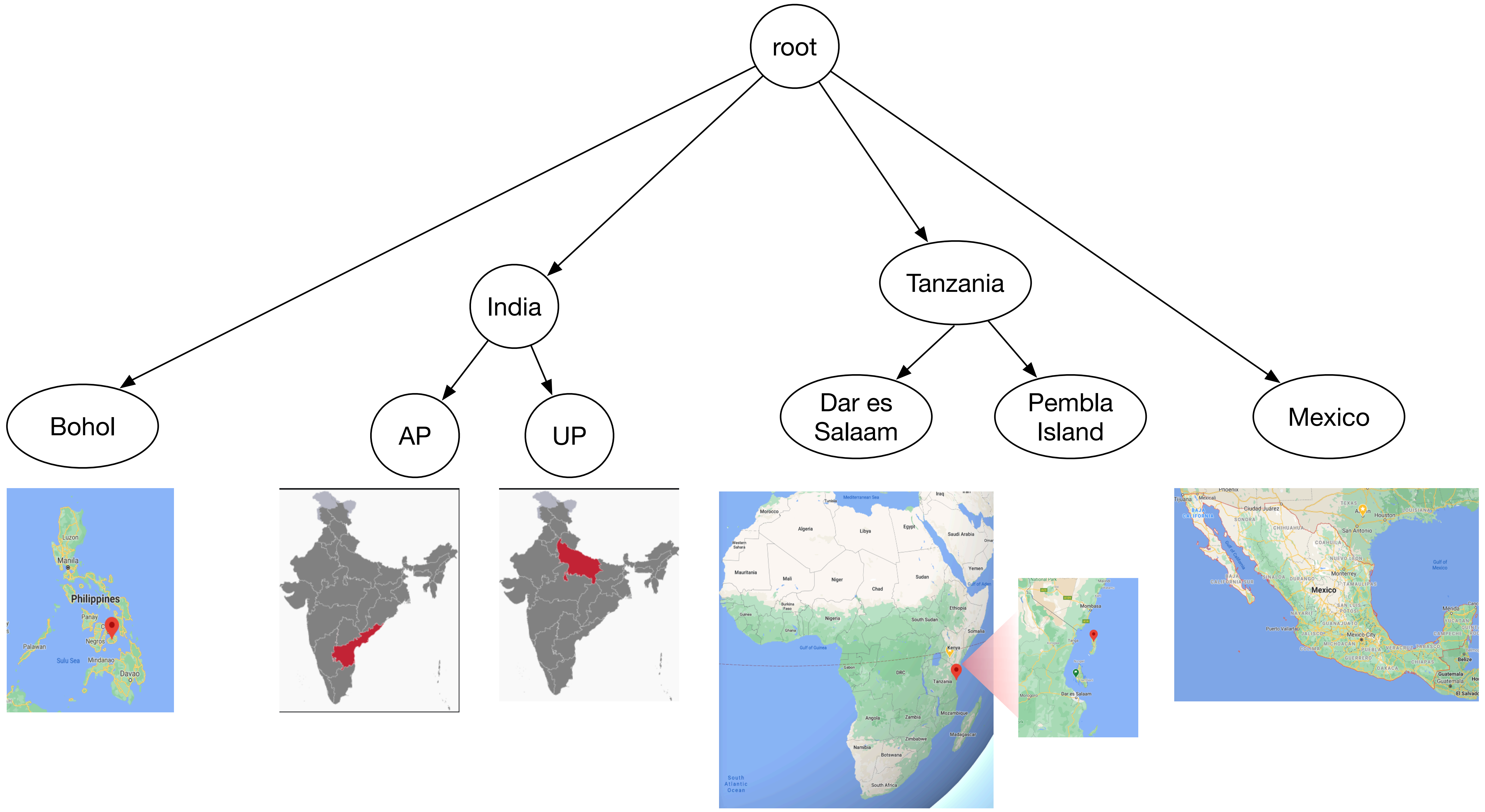
- Expansion of VA to new “domains": new regions (e.g., Brazil, New Guinea) and/or new time periods (COVID vs non-COVID periods): $$ Gates Foundation/ Bloomberg Philanthropies
  - potential data distribution shifts call for domain-adaptive methods
  - New statistical question:
    - Can we estimate cause-specific mortality fractions (CSMFs) with some robustness to data distribution shifts between the source and the target domains?

Population Health Metrics Research Consortium (PHMRC) Verbal Autopsy Data

- The PHMRC VA gold-standard data (Population Health Metrics Research Consortium, 2018):
  - Mexico City, Mexico
  - Andhra Pradesh, India
  - Uttar Pradesh, India
  - Dar es Salaam, Tanzania
  - Pemba Island, Tanzania
  - Bohol, Philippines.

- Gold-standard CODs are obtained from clinical diagnostics.

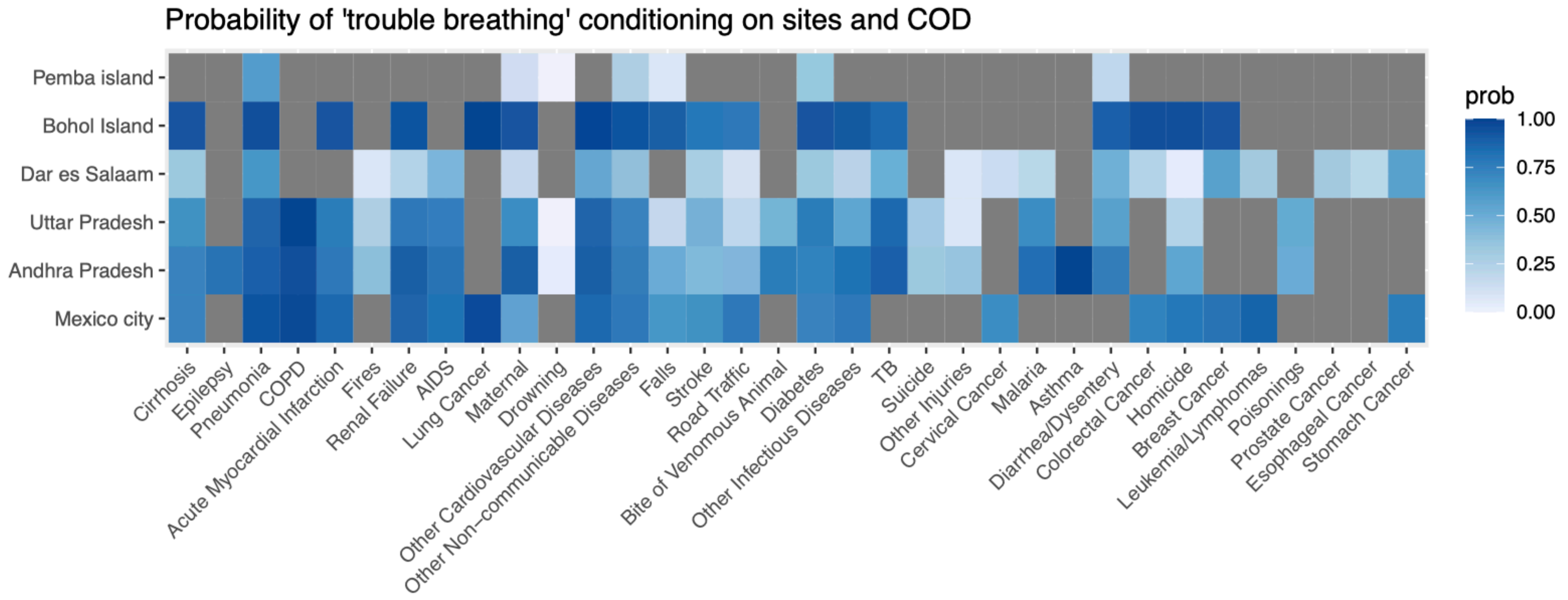
- We will take one site as the target and use the other five sites as source domains.

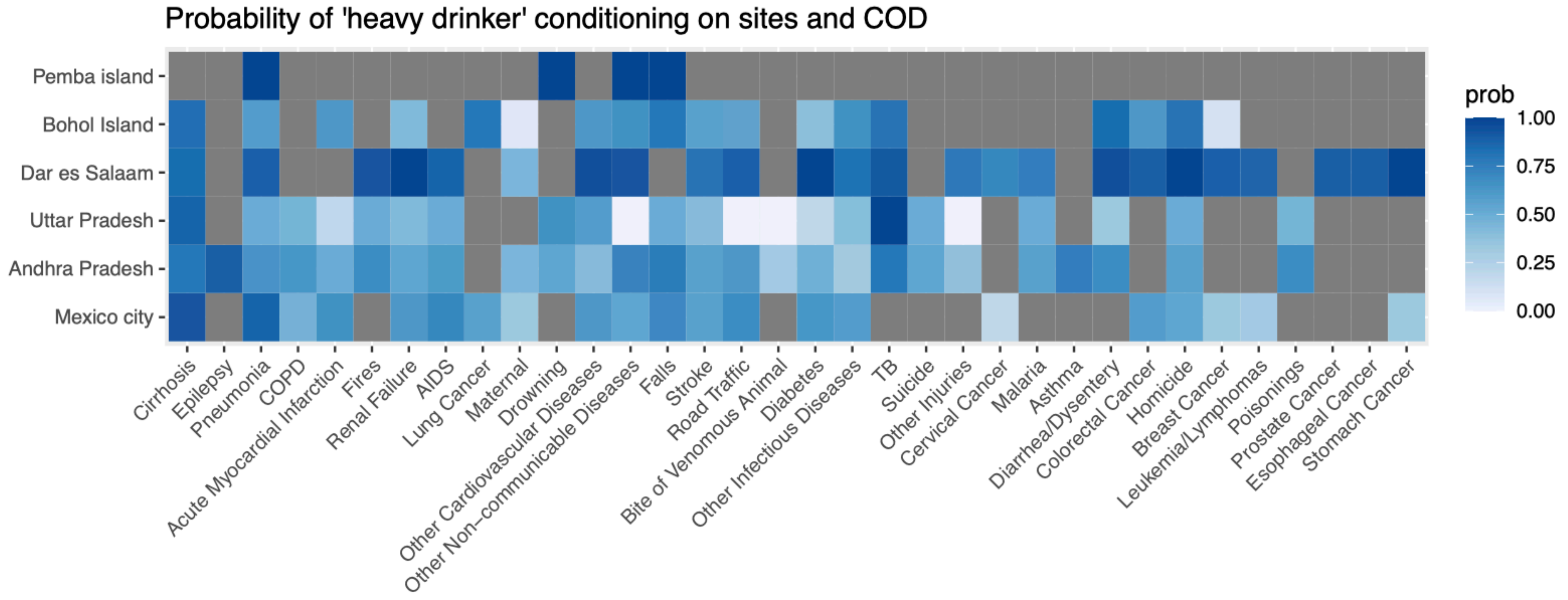


Hierarchy encodes geographical similarity

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.

Example of Between-Domain Differences: PHMRC Data


Probability of 'trouble breathing' conditioning on sites and COD


Probability of 'heavy drinker' conditioning on sites and COD

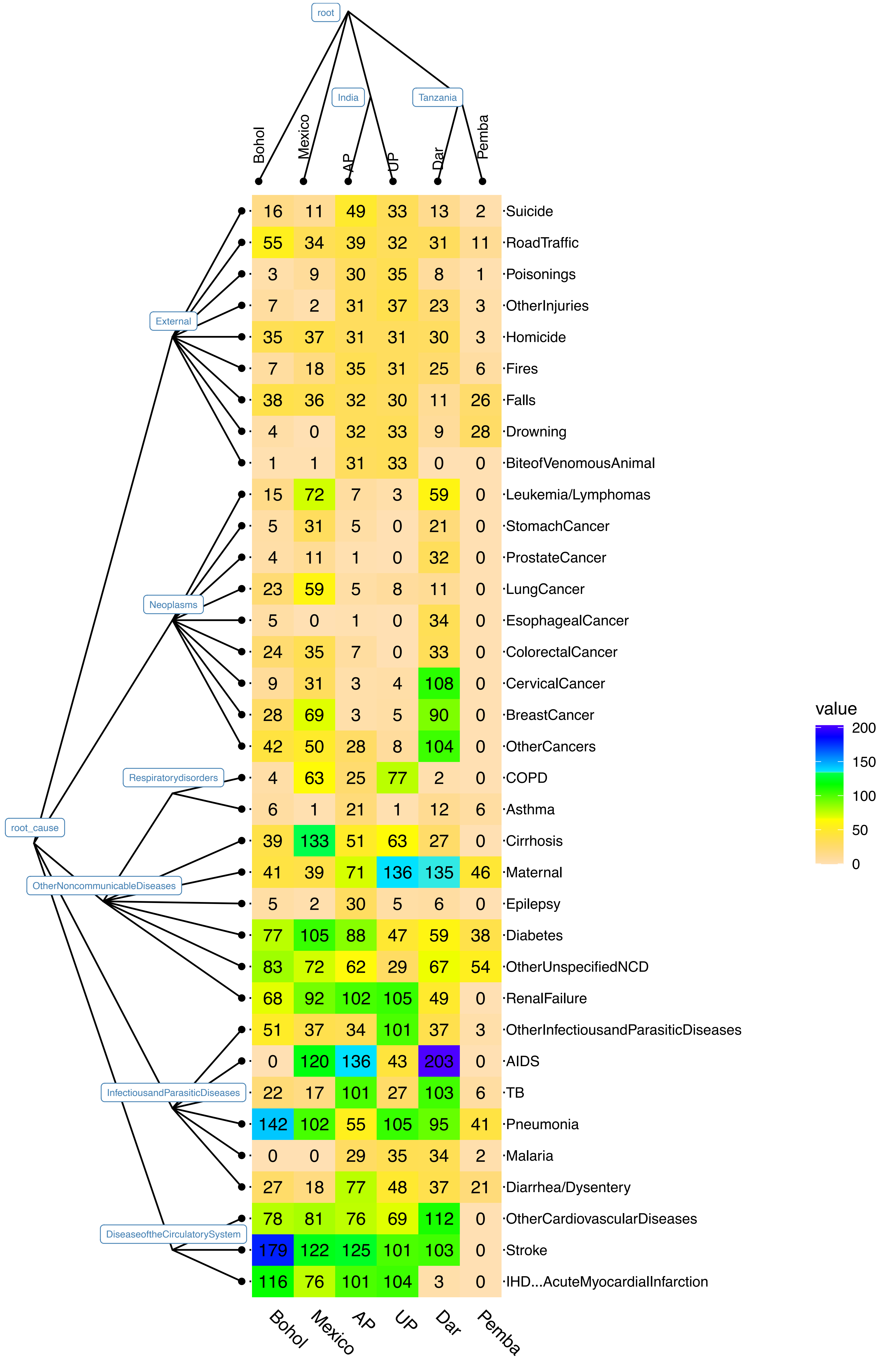
Plots including only symptom–cause pairs with least 20 observations.

Data: A Closer Look

Death Counts

Death counts by 35 causes and 6 sites 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.

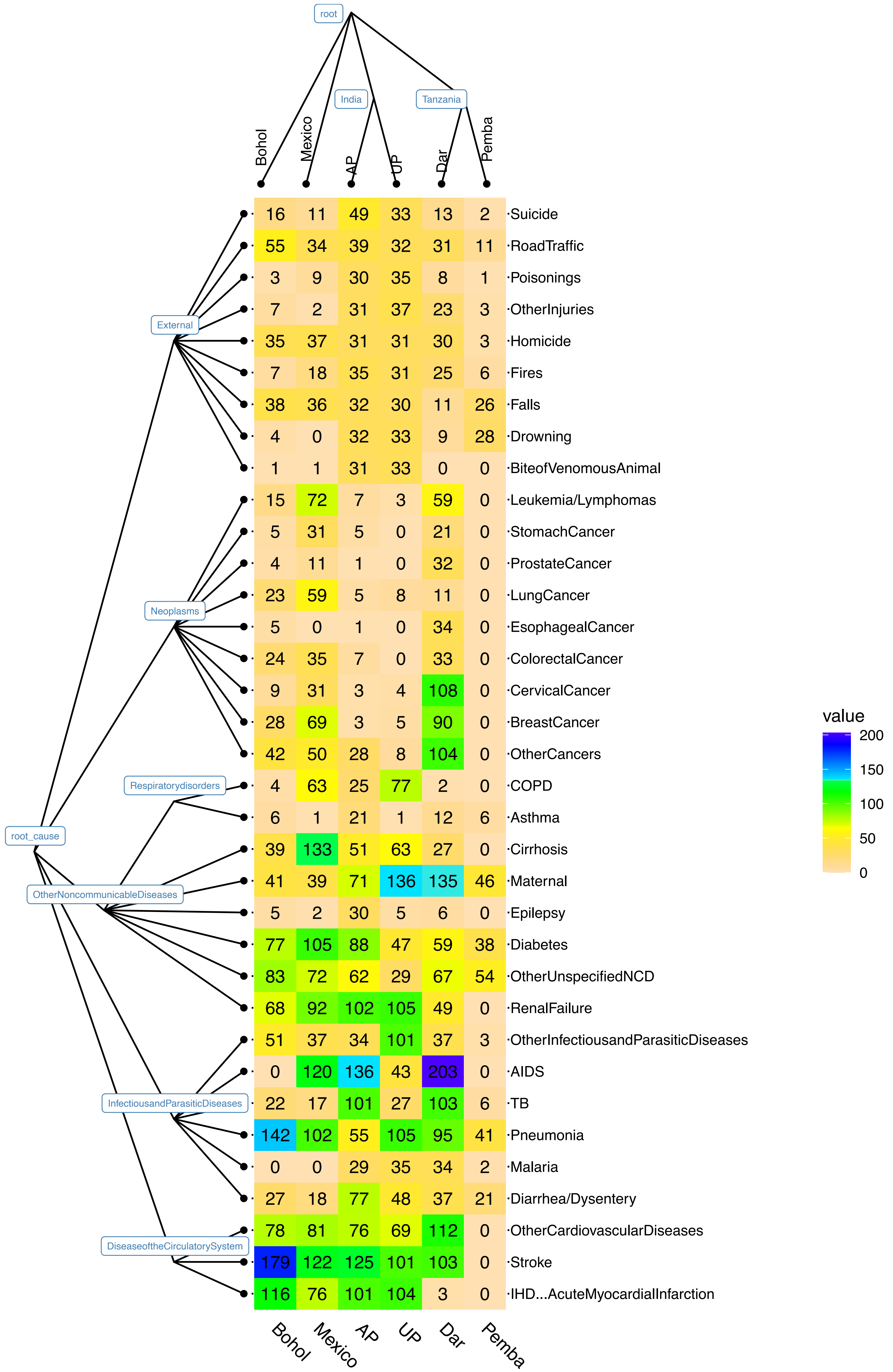


	Bohol	Mexico	AP	UP	Dar	Pemba
Suicide	16	11	49	33	13	2
RoadTraffic	55	34	39	32	31	11
Poisonings	3	9	30	35	8	1
OtherInjuries	7	2	31	37	23	3
Homicide	35	37	31	31	30	3
Fires	7	18	35	31	25	6
Falls	38	36	32	30	11	26
Drowning	4	0	32	33	9	28
BiteofVenomousAnimal	1	1	31	33	0	0
Leukemia/Lymphomas	15	72	7	3	59	0
StomachCancer	5	31	5	0	21	0
ProstateCancer	4	11	1	0	32	0
LungCancer	23	59	5	8	11	0
EsophagealCancer	5	0	1	0	34	0
ColorectalCancer	24	35	7	0	33	0
CervicalCancer	9	31	3	4	108	0
BreastCancer	28	69	3	5	90	0
OtherCancers	42	50	28	8	104	0
COPD	4	63	25	77	2	0
Asthma	6	1	21	1	12	6
Cirrhosis	39	133	51	63	27	0
Maternal	41	39	71	136	135	46
Epilepsy	5	2	30	5	6	0
Diabetes	77	105	88	47	59	38
OtherUnspecifiedNCD	83	72	62	29	67	54
RenalFailure	68	92	102	105	49	0
OtherInfectiousandParasiticDiseases	51	37	34	101	37	3
AIDS	0	120	136	43	203	0
TB	22	17	101	27	103	6
Pneumonia	142	102	55	105	95	41
Malaria	0	0	29	35	34	2
Diarrhea/Dysentery	27	18	77	48	37	21
OtherCardiovascularDiseases	78	81	76	69	112	0
Stroke	179	122	125	101	103	0
IHD...AcuteMyocardialInfarction	116	76	101	104	3	0

Data: A Closer Look

Death Counts

- Sparse table
- “Small area estimation”
- Trees provide prior information about similarities between the domains (among the columns)
- We have assumed trees are given



	Bohol	Mexico	AP	UP	Dar	Pemba
Suicide	16	11	49	33	13	2
RoadTraffic	55	34	39	32	31	11
Poisonings	3	9	30	35	8	1
OtherInjuries	7	2	31	37	23	3
Homicide	35	37	31	31	30	3
Fires	7	18	35	31	25	6
Falls	38	36	32	30	11	26
Drowning	4	0	32	33	9	28
BiteofVenomousAnimal	1	1	31	33	0	0
Leukemia/Lymphomas	15	72	7	3	59	0
StomachCancer	5	31	5	0	21	0
ProstateCancer	4	11	1	0	32	0
LungCancer	23	59	5	8	11	0
EsophagealCancer	5	0	1	0	34	0
ColorectalCancer	24	35	7	0	33	0
CervicalCancer	9	31	3	4	108	0
BreastCancer	28	69	3	5	90	0
OtherCancers	42	50	28	8	104	0
COPD	4	63	25	77	2	0
Asthma	6	1	21	1	12	6
Cirrhosis	39	133	51	63	27	0
Maternal	41	39	71	136	135	46
Epilepsy	5	2	30	5	6	0
Diabetes	77	105	88	47	59	38
OtherUnspecifiedNCD	83	72	62	29	67	54
RenalFailure	68	92	102	105	49	0
OtherInfectiousandParasiticDiseases	51	37	34	101	37	3
AIDS	0	120	136	43	203	0
TB	22	17	101	27	103	6
Pneumonia	142	102	55	105	95	41
Malaria	0	0	29	35	34	2
Diarrhea/Dysentery	27	18	77	48	37	21
OtherCardiovascularDiseases	78	81	76	69	112	0
Stroke	179	122	125	101	103	0
IHD...AcuteMyocardialInfarction	116	76	101	104	3	0

Data: A Closer Look

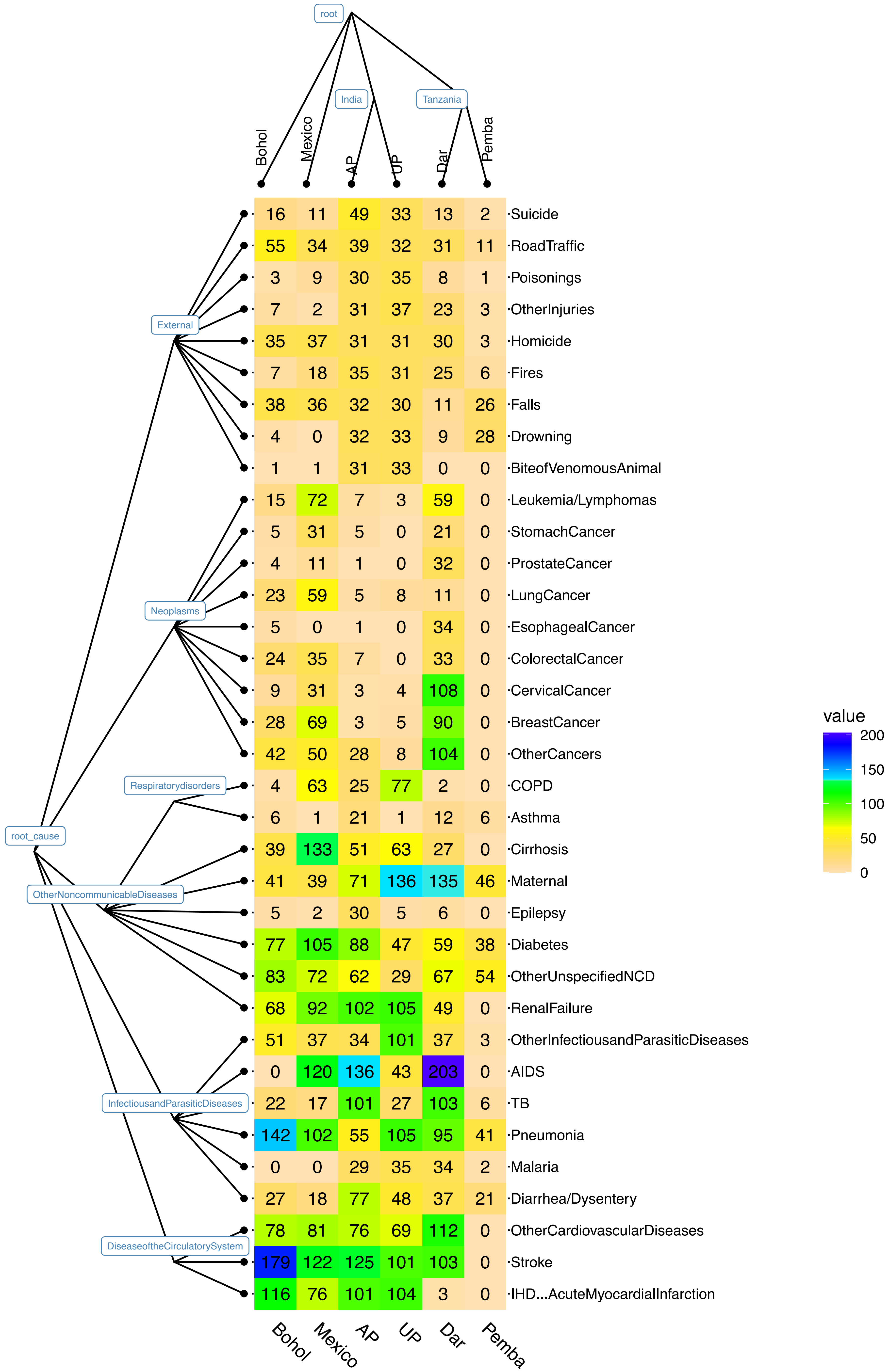
Death Counts

We will mask the causes-of-death in one site (column) 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 site hierarchy are *a priori* more likely fused to have the same conditional distribution of the VA survey responses given a cause

- we will allow the degree of similarity may vary by cause (“drowning” vs “”)



Bohol	Mexico	AP	UP	Dar	Pemba	
16	11	49	33	13	2	Suicide
55	34	39	32	31	11	RoadTraffic
3	9	30	35	8	1	Poisonings
7	2	31	37	23	3	OtherInjuries
35	37	31	31	30	3	Homicide
7	18	35	31	25	6	Fires
38	36	32	30	11	26	Falls
4	0	32	33	9	28	Drowning
1	1	31	33	0	0	BiteofVenomousAnimal
15	72	7	3	59	0	Leukemia/Lymphomas
5	31	5	0	21	0	StomachCancer
4	11	1	0	32	0	ProstateCancer
23	59	5	8	11	0	LungCancer
5	0	1	0	34	0	EsophagealCancer
24	35	7	0	33	0	ColorectalCancer
9	31	3	4	108	0	CervicalCancer
28	69	3	5	90	0	BreastCancer
42	50	28	8	104	0	OtherCancers
4	63	25	77	2	0	COPD
6	1	21	1	12	6	Asthma
39	133	51	63	27	0	Cirrhosis
41	39	71	136	135	46	Maternal
5	2	30	5	6	0	Epilepsy
77	105	88	47	59	38	Diabetes
83	72	62	29	67	54	OtherUnspecifiedNCD
68	92	102	105	49	0	RenalFailure
51	37	34	101	37	3	OtherInfectiousandParasiticDiseases
0	120	136	43	203	0	AIDS
22	17	101	27	103	6	TB
142	102	55	105	95	41	Pneumonia
0	0	29	35	34	2	Malaria
27	18	77	48	37	21	Diarrhea/Dysentery
78	81	76	69	112	0	OtherCardiovascularDiseases
179	122	125	101	103	0	Stroke
116	76	101	104	3	0	IHD...AcuteMyocardialInfarction

Notation

- $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iJ})^{\mathsf{T}} \in \{0,1\}^J$: a vector of binary responses for subject i = 1, …, N
- (Y_i, D_i): (cause of death, domain)
  - Y_i takes value from $\{1, \ldots, C\}$, indicating the cause of death among a total of pre-specified causes
  - D_i takes its value from $\{0, 1, \ldots, G\}$, indicating domain membership: 0 for target domain, 1 to G for the G pre-specified source domains
- D_i is assumed to be observed for all subjects
- Y_i observed for $\{i : D_i \neq 0\}$ in the source domains; unobserved otherwise

Notation

- 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.

Distribution Shift in VA Data

- For a domain, the joint distribution of (causes of death, VA responses) can be factored into

  - a) a vector of population-level marginal probabilities of the causes (or “cause-specific mortality fractions”, CSMF)
  - b) conditional distribution of the VA responses given a cause

- a) CSMF may differ by domain: most natural - a cause may differentially contribute to deaths occurred in different study populations.
- b) may differ by domain
  - Need “intelligent information pooling between the domains”

Statistical Considerations

- a) a vector of population-level marginal probabilities of the causes (or “cause-specific mortality fractions”, CSMF)
- b) conditional distribution of the VA responses given a cause

- Better estimation of (b) → better estimation of (a) (Kunihama et al. 2020; Li, McCormick and Clark et al. 2020)
- Estimate (b) domain by domain?
  - Pros: recognize between-domain differences in (b)
  - Cons: certain domains may have few sampled deaths due to a cause → statistical instability in estimating (b) → inaccurate estimation of (a)

Statistical Considerations

- a) a vector of population-level marginal probabilities of the causes (or “cause-specific mortality fractions”, CSMF)
- b) conditional distribution of the VA responses given a cause

- Solution: pooling information between domains to estimation (b)s for the domains
  - “complete pooling”: force domains to have the same (b) → restrictive and obscures potential between-domain differences in (b) 🫠
  - “ad hoc partial pooling”: potentially mis-specified grouping of domains for pooled estimation of (b) 🥲
  - We will do “data-driven pooling, guided by a domain hierarchy (‘tree’)” 😇

Existing Literature

- Datta et al. (2021), Fiksel et al. (2021): Methods that calibrate CSMF estimates obtained from VA algorithms trained on a training data set to produce CSMF estimates in a new population
  - Differ from our work in three important ways
    - Only estimated CSMFs from a list of trained VA algorithms; not designed for individual-level information in the training data to perform calibration
    - Relies on a small # of deaths with medically-confirmed causes in the new population
    - Causes often need to be manually combined to produce stable and meaningful results
- Moran et al. (2021):
  - Factor regression to let conditional distributions of symptoms given a cause depend on individual-level covariates, including domain dummy variables
  - But not designed for a domain with no observed CODs

Our Framework: Nested Latent Class Models

We assume the following model specifications for $\mathcal{D}$:

$$\begin{aligned} \text{cause of death}: &\ Y_i \mid D_i = g \sim \text{Categorical}_C(\boldsymbol{\pi}^{(g)}), & (1)\\ \text{latent class}: &\ Z_i \mid Y_i = c, D_i = g \sim \text{Categorical}_K(\boldsymbol{\lambda}^{(c,g)}), & (2)\\ \text{responses}: &\ X_{ij} \mid Z_i = k, Y_i = c \overset{\text{indep.}}{\sim} \text{Bernoulli}(\theta_{jk}^{(c)}), j \in [J] & (3) \end{aligned}$$

for $i \in [N]$, $g \in \{0\} \cup [G]$, where the population parameters $\boldsymbol{\pi}^{(g)} = (\pi_1^{(g)}, \ldots, \pi_C^{(g)})^\intercal$ 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 $\{Y_i : D_i = 0\}$.

Why bother?

- Given each cause, the conditional distribution of symptom approximated by a latent class model (read “parallel factor decomposition”). Relative to Gaussian thresholded approaches
    - easy to control the number of classes, to induce parsimony
    - computationally (much) easier
- New domain having new “innovations in the symptom distributions”?
    - Add additional classes

Prior Distribution to Integrate the Tree Information

Condensed Summary

- Tree-informed Bayesian shrinkage prior
  - a general framework: Gaussian diffusion process for parameters realized from the root to the leaves of a tree, then transform to parameters that enter the observed likelihood via, e.g., expit(), i.e., “sigmoid”
  - heuristics: “parameters connected by shorter paths in a tree *a priori* take more similar values”
  - we apply this framework to let two distinct sets of parameters diffuse along respective trees (domain hierarchy - for lambda’s, cause hierarchy -for theta’s)

Desired Outputs

- Primary Outputs
  - Population level (the focus of this talk)
    - Cause Specific Mortality Fractions (CSMF)
      - posterior mean, posterior distribution
  - Individual level
    - Probability of COD for each death
      - a vector of probabilities that assign a death to all pre-specified categories of CODs

Desired Outputs II

- Secondary Outputs
    - Data-driven discovery of leaf groups in the domain hierarchies
        - selectively collapse along the trees
        - parameter fusion: “which domains’ data can be pooled for estimation as informed by their similarities from data”
        - Similarity measure between the leaves (cophenetic distance)

Nested Latent Class Models

Variational Algorithm for Approximate Posterior Inference

1. We use *variational Bayes* to conduct approximate posterior inference (Blei, Kucukelbir and Mcauliffe, 2017; Thomas et al. 2019)

2. This is more scalable for large trees and large sample sizes

3. This overcomes some known sampling issues with MCMC for dealing spike-and-slab priors (George and McCulloch, 1997)

R package 🎄🎄: https://github.com/zhenkewu/doubletree

The package is designed to work under all possible patterns of observed and missing causes of death

Simulation Design

- Setup: G training domains (g = 1, 2, …, G), 1 target domain (g=0)
- Simulate VA response data and true CODs for all domains according to the true model
- Choose one domain as “target”, mask all or a subset of the chosen domain’s CODs
    - Scenarios where the proposed model will likely compare favorably to alternatives
    - Scenarios where the proposed model will be more or less similar to alternatives

Results

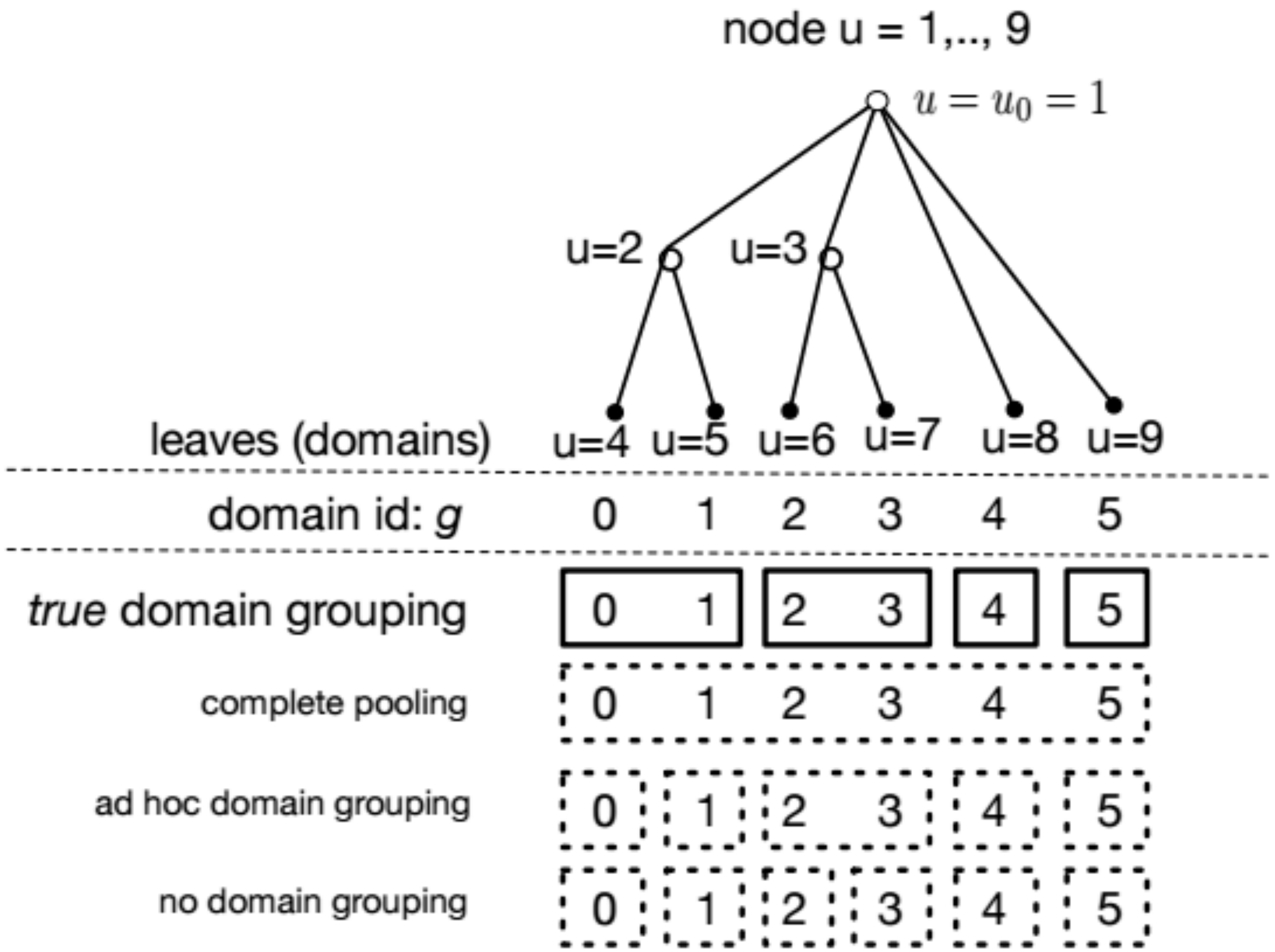
- Performance Metrics
  - CSMF accuracy: normalized L1 distance

$$ACC_{csmf} = 1 - \frac{\sum_{c=1}^{C} |CSMF_c^{true} - CSMF_c^{pred}|}{2(1 - \min CSMF^{true})}$$

- Top cause accuracy, top 3 cause accuracy (McCormick et al. 2016 JASA)

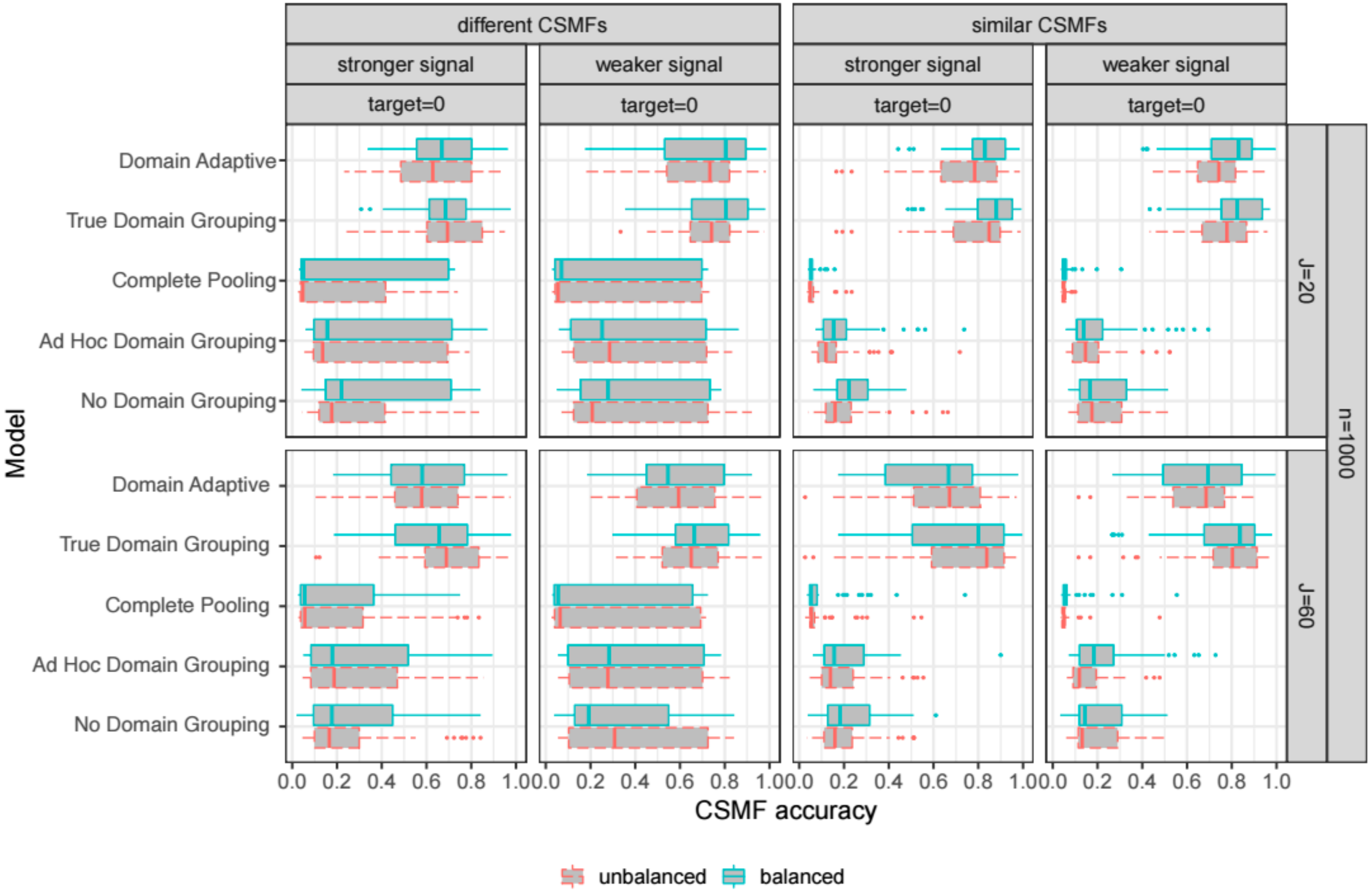
- CSMF accuracy (Murray et al. 2011, Population Health Metrics)

Simulation Results



(a) Simulation I: domain tree and different domain groupings used in comparison.

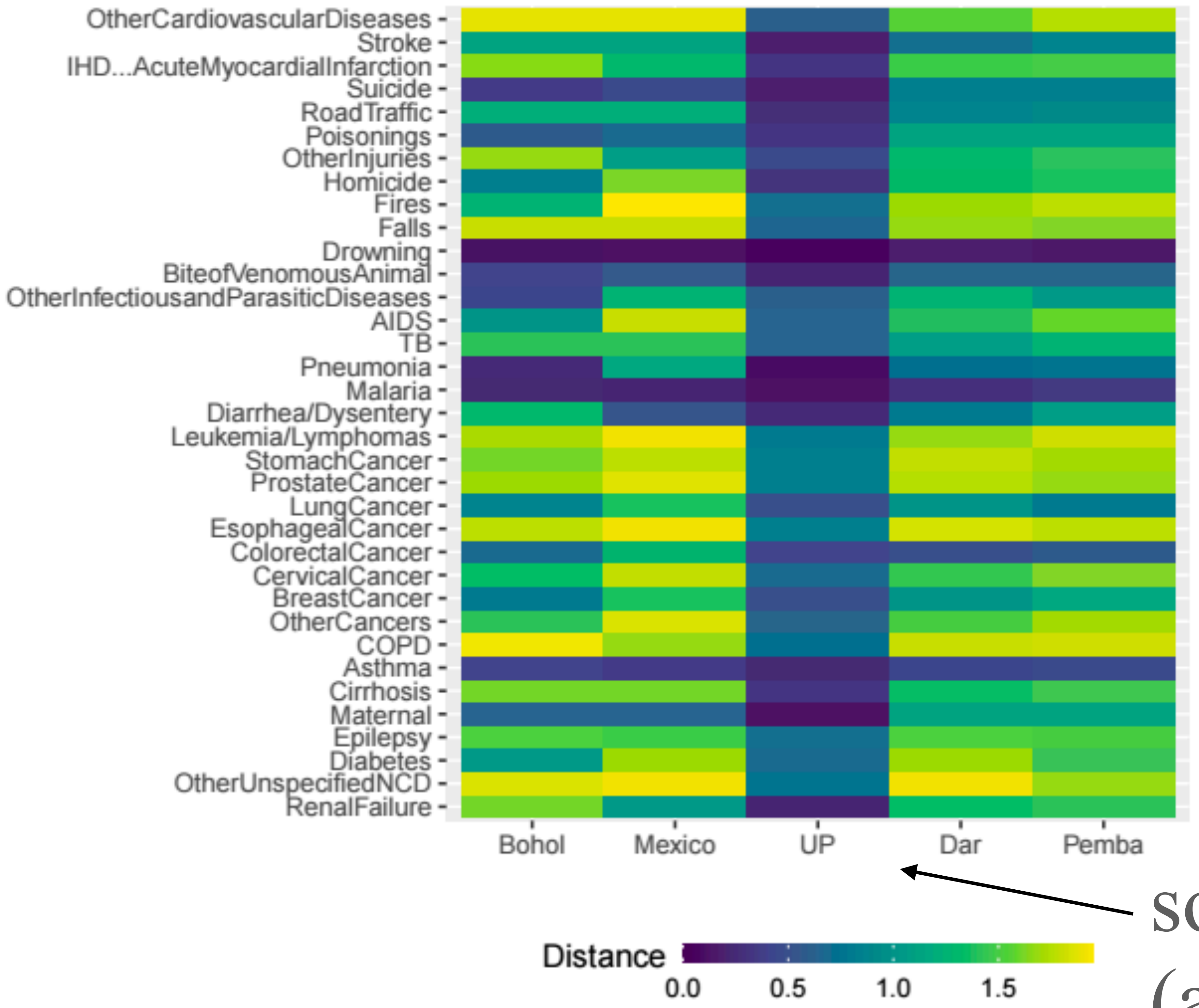
Simulation Results



(b) Simulation I: CSMF accuracy comparison.

PHMRC Data Results: “Source-Target Similarity”

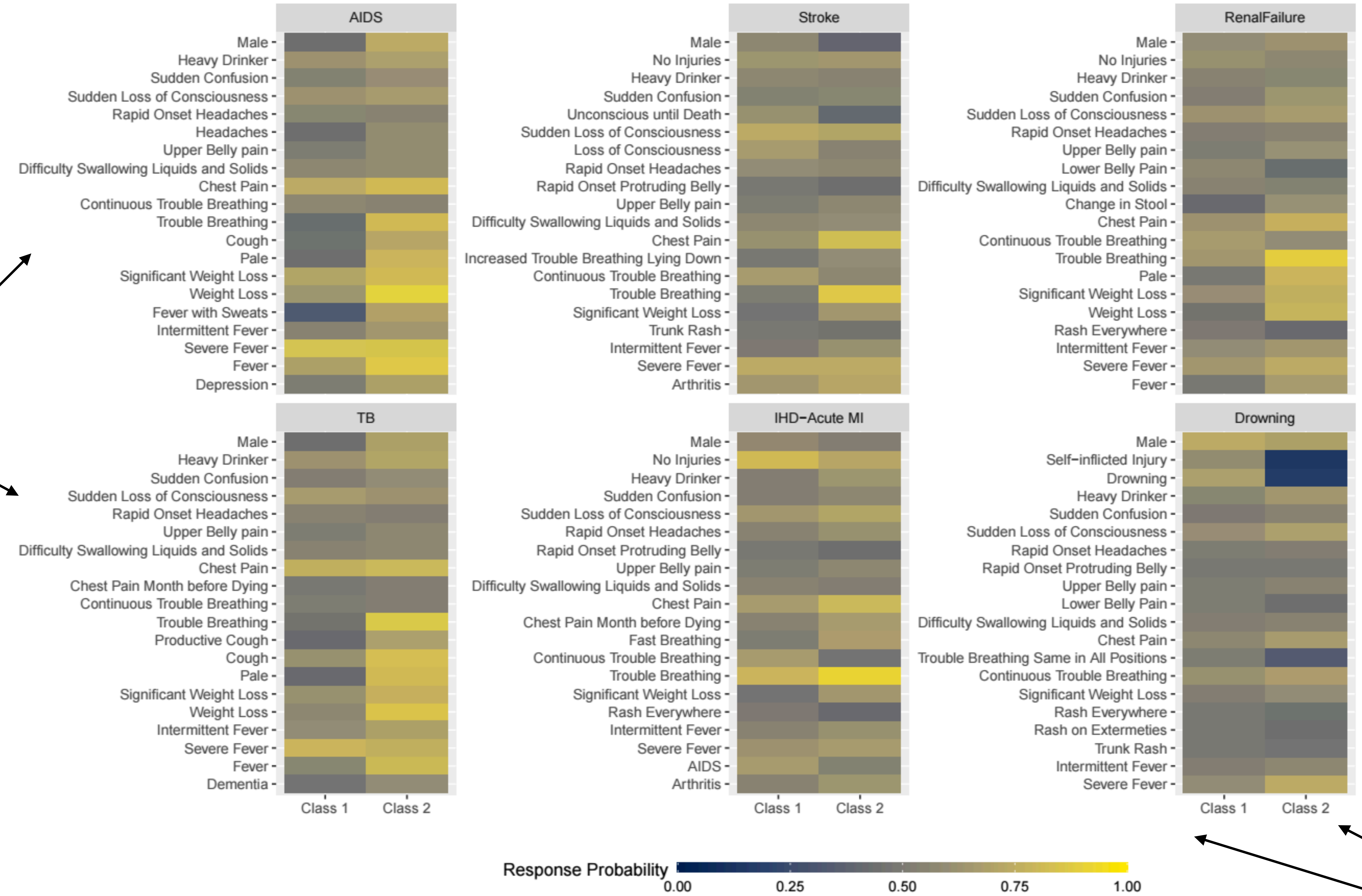
causes



source domains
(all but the target domain
“AP”, a site from India)

(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.

PHMRC Data Results: “class profiles”



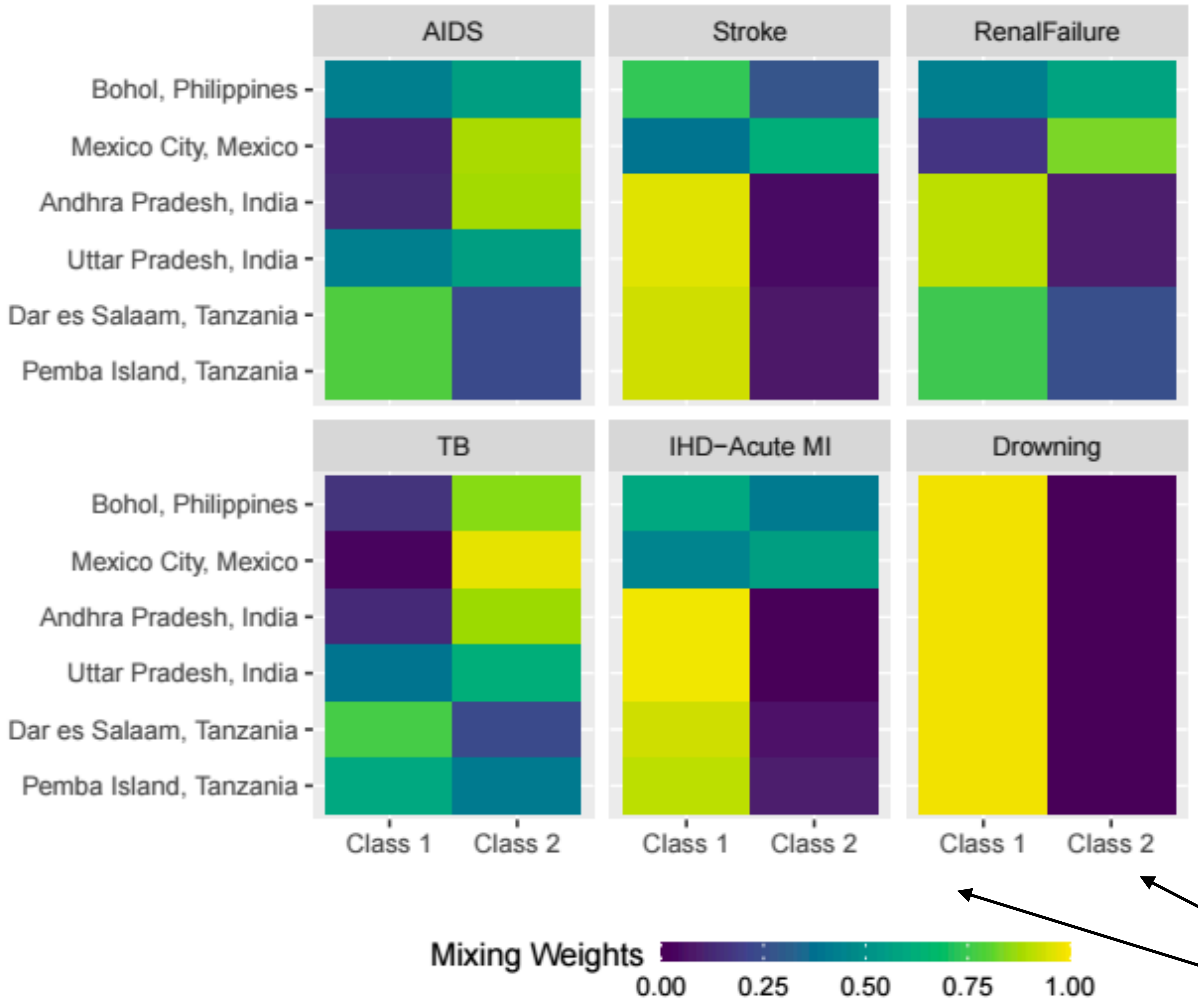
causes

surveyed “symptoms”

classes

(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).

PHMRC Data Results: “Domains differ in how the classes got mixed”



causes

domains

classes

(b) Variation of class-mixing weights between domains; six sets of weights are shown for six causes of deaths (the model uses 35 causes).

Main Points Once Again

- Distribution shifts between the source and target domains are common, e.g.,
  - In VA, conditional distributions of symptoms given a cause may vary by study sites
  - The degree of this variation may differ by cause
- Domain adaptive method is needed for improving the estimation of the target domain's population-level parameters and individual-level predictions
- Among many possible solutions, the present work focused on
  - "how to use a tree to guide domain adaptation?"
- For illustration, we used a domain tree that encodes geographic similarity information.
  - One can use domain-level info to form a hierarchy, e.g., by hierarchical clustering, and then use that tree as input for our method

Future Directions

- Method directions:
  - Same set of response probability profiles? Can be relaxed using techniques from recent robust clustering work (Stephenson et al. (2020))
  - Different K’s across causes
  - General graph-informed clustering with tensor decomposition approximation
  - Negative transfer issues: “a bad module/additional noisy data may harm statistical performances. This has been noted in Multitask Gaussian Process literature.
- Applied directions:
  - How to deal with emerging prominent causes over different time periods (COVID19…)?
  - COD labels might be noisy:
    - How to do privacy-robust analysis (“Died of Malaria, but in fact…”)? Adversarial-labeling resistant analysis?

Paper:

Wu et al. (2022+). Tree-informed Bayesian multi-source domain adaptation: cross-population probabilistic cause-of-death assignment using verbal autopsy. https://arxiv.org/abs/2112.10978

Software:

R package 🎄🎄: https://github.com/zhenkewu/doubletree

The package is designed to work under all possible patterns of observed and missing causes of death

Thank you!

zhenkewu@umich.edu