Statistical methods for mortality estimation in data-sparse settings

Instructions: Click on the link to access each author's presentation.

Organiser: Zehang Li

Chair: Tyler McCormick

Participants:

Zehang Li: Subpopulation mortality surveillance using verbal autopsies

Monica Alexander: Estimating the timing of stillbirths worldwide

Myriam Cifuentes:* The role of demographics of age in COVID contact tracing and contagion networks

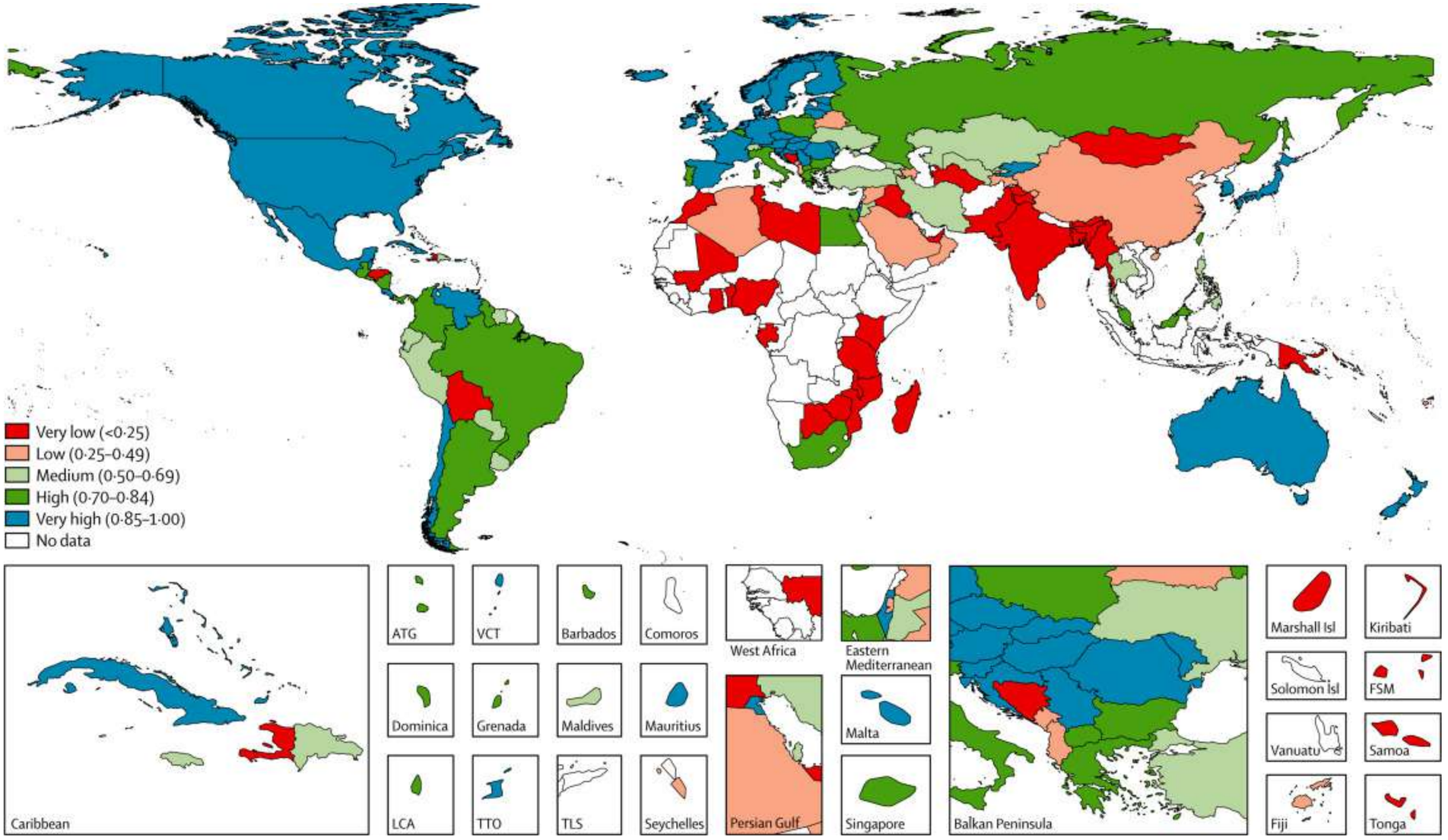
Zhenke Wu: Enhancing Mortality Estimation: Cooperative Distribution-Valued Matrix Completion to Integrate Expert Prior Knowledge

*Work presentation not available or non-existent

Statistical Methods for Mortality Estimation in Data-Sparse Settings

Zehang Richard Li
Department of Statistics
University of California, Santa Cruz

IAOS-ISI
May 15, 2024



"Globally, only modest progress has been made since 2000, with the percentage of deaths registered increasing from 36% to 38%, and the percentage of children aged under 5 years whose birth has been registered increasing from 58% to 65%."

Mikkelsen et al., Lancet, 2015

Overview

Overview

- Monica Alexander talked about estimating **proportion of intrapartum stillbirths**

Overview

- Monica Alexander talked about estimating **proportion of intrapartum stillbirths**
    - Data: CRVS, health management information systems, health facility, Global Network Study, UN IGME estimates of NMR, SFB, etc.

Overview

- Monica Alexander talked about estimating **proportion of intrapartum stillbirths**
  - Data: CRVS, health management information systems, health facility, Global Network Study, UN IGME estimates of NMR, SFB, etc.
  - Challenges: Different data quality, missing proportion, and measurement errors across sources.

Overview

- Monica Alexander talked about estimating **proportion of intrapartum stillbirths**
  - Data: CRVS, health management information systems, health facility, Global Network Study, UN IGME estimates of NMR, SFB, etc.
  - Challenges: Different data quality, missing proportion, and measurement errors across sources.
  - Method: Pooling data from all countries in a regression model and model the shared components.

Overview

- Monica Alexander talked about estimating **proportion of intrapartum stillbirths**
  - Data: CRVS, health management information systems, health facility, Global Network Study, UN IGME estimates of NMR, SFB, etc.
  - Challenges: Different data quality, missing proportion, and measurement errors across sources.
  - Method: Pooling data from all countries in a regression model and model the shared components.
- Zhenke Wu talked about estimating **cause-specific mortality fractions**

Overview

- Monica Alexander talked about estimating **proportion of intrapartum stillbirths**
  - Data: CRVS, health management information systems, health facility, Global Network Study, UN IGME estimates of NMR, SFB, etc.
  - Challenges: Different data quality, missing proportion, and measurement errors across sources.
  - Method: Pooling data from all countries in a regression model and model the shared components.
- Zhenke Wu talked about estimating **cause-specific mortality fractions**
  - Data: Verbal autopsy from multiple non-local populations.

Overview

- Monica Alexander talked about estimating **proportion of intrapartum stillbirths**
  - Data: CRVS, health management information systems, health facility, Global Network Study, UN IGME estimates of NMR, SFB, etc.
  - Challenges: Different data quality, missing proportion, and measurement errors across sources.
  - Method: Pooling data from all countries in a regression model and model the shared components.
- Zhenke Wu talked about estimating **cause-specific mortality fractions**
  - Data: Verbal autopsy from multiple non-local populations.
  - Challenges: Relationship between symptoms and causes change over populations.

Overview

- Monica Alexander talked about estimating **proportion of intrapartum stillbirths**
  - Data: CRVS, health management information systems, health facility, Global Network Study, UN IGME estimates of NMR, SFB, etc.
  - Challenges: Different data quality, missing proportion, and measurement errors across sources.
  - Method: Pooling data from all countries in a regression model and model the shared components.
- Zhenke Wu talked about estimating **cause-specific mortality fractions**
  - Data: Verbal autopsy from multiple non-local populations.
  - Challenges: Relationship between symptoms and causes change over populations.
  - Method: Pooling data from multiple populations and model the shared components and heterogeneity.

Discussion overview

Discussion overview

- A key theme in both talks is how to combine information from multiple datasets.

Discussion overview

- A key theme in both talks is how to combine information from multiple datasets.

- More specifically, how to combine *weak* information…

Discussion overview

- A key theme in both talks is how to combine information from multiple datasets.
- More specifically, how to combine *weak* information…
- This is an important problem in many population health research domains.

Discussion overview

- A key theme in both talks is how to combine information from multiple datasets.
- More specifically, how to combine *weak* information…
- This is an important problem in many population health research domains.
- I will do a brief recap of both talks, mention some related topics (from my work), and mention some questions and thoughts.

Estimating intrapartum stillbirths (IPSB) fraction

Estimating intrapartum stillbirths (IPSB) fraction

- The paper jointly models different data sources over 92 countries!

Estimating intrapartum stillbirths (IPSB) fraction

- The paper jointly models different data sources over 92 countries!
- Place-level prevalence is decomposed into

Estimating intrapartum stillbirths (IPSB) fraction

- The paper jointly models different data sources over 92 countries!
- Place-level prevalence is decomposed into
  - Main effects from region, country, subpopulation,

Estimating intrapartum stillbirths (IPSB) fraction

- The paper jointly models different data sources over 92 countries!
- Place-level prevalence is decomposed into
  - Main effects from region, country, subpopulation,
  - Fixed effect from NMR (log scale),

Estimating intrapartum stillbirths (IPSB) fraction

- The paper jointly models different data sources over 92 countries!
- Place-level prevalence is decomposed into
  - Main effects from region, country, subpopulation,
  - Fixed effect from NMR (log scale),
  - Place-specific time trends,

Estimating intrapartum stillbirths (IPSB) fraction

- The paper jointly models different data sources over 92 countries!
- Place-level prevalence is decomposed into
  - Main effects from region, country, subpopulation,
  - Fixed effect from NMR (log scale),
  - Place-specific time trends,
  - Adjustment for gestational age definition,

Estimating intrapartum stillbirths (IPSB) fraction

- The paper jointly models different data sources over 92 countries!
- Place-level prevalence is decomposed into
  - Main effects from region, country, subpopulation,
  - Fixed effect from NMR (log scale),
  - Place-specific time trends,
  - Adjustment for gestational age definition,
  - Additional noise based on study type, if not CRVS.

Estimating intrapartum stillbirths (IPSB) fraction

- The paper jointly models different data sources over 92 countries!
- Place-level prevalence is decomposed into
  - Main effects from region, country, subpopulation,
  - Fixed effect from NMR (log scale),
  - Place-specific time trends,
  - Adjustment for gestational age definition,
  - Additional noise based on study type, if not CRVS.
- Country-level prevalence are estimated as weighted average of place-level prevalence.

Estimating intrapartum stillbirths (IPSB) fraction

- The paper jointly models different data sources over 92 countries!
- Place-level prevalence is decomposed into
  - Main effects from region, country, subpopulation,
  - Fixed effect from NMR (log scale),
  - Place-specific time trends,
  - Adjustment for gestational age definition,
  - Additional noise based on study type, if not CRVS.
- Country-level prevalence are estimated as weighted average of place-level prevalence.
  - Weights computed by comparing observed counts with UN IGME estimates

Estimating intrapartum stillbirths (IPSB) fraction

Estimating intrapartum stillbirths (IPSB) fraction

- The model is constructed very carefully with lots of thoughts going into the model component, specification, and how to properly account for data quality.

Estimating intrapartum stillbirths (IPSB) fraction

- The model is constructed very carefully with lots of thoughts going into the model component, specification, and how to properly account for data quality.

- Some similar efforts in modeling key demographic indicators:

Estimating intrapartum stillbirths (IPSB) fraction

- The model is constructed very carefully with lots of thoughts going into the model component, specification, and how to properly account for data quality.
- Some similar efforts in modeling key demographic indicators:

The Annals of Applied Statistics
2014, Vol. 8, No. 4, 2122–2149
DOI: 10.1214/14-AOAS768
© Institute of Mathematical Statistics, 2014

GLOBAL ESTIMATION OF CHILD MORTALITY USING A BAYESIAN B-SPLINE BIAS-REDUCTION MODEL[1]

BY LEONTINE ALKEMA AND JIN ROU NEW

National University of Singapore

Estimating intrapartum stillbirths (IPSB) fraction

- The model is constructed very carefully with lots of thoughts going into the model component, specification, and how to properly account for data quality.
- Some similar efforts in modeling key demographic indicators:

The Annals of Applied Statistics
2014, Vol. 8, No. 4, 2122–2149
DOI: 10.1214/14-AOAS768
© Institute of Mathematical Statistics, 2014

GLOBAL ESTIMATION OF CHILD MORTALITY USING A BAYESIAN B-SPLINE BIAS-REDUCTION MODEL[1]

BY LEONTINE ALKEMA AND JIN ROU NEW

National University of Singapore

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1245–1274
DOI: 10.1214/16-AOAS1014
© Institute of Mathematical Statistics, 2017

A BAYESIAN APPROACH TO THE GLOBAL ESTIMATION OF MATERNAL MORTALITY[1]

BY LEONTINE ALKEMA*, SANQIAN ZHANG†, DORIS CHOU‡, ALISON GEMMILL§, ANN-BETH MOLLER‡, DORIS MA FAT‡, LALE SAY‡, COLIN MATHERS‡ AND DANIEL HOGAN‡

University of Massachusetts, Amherst, Harvard University†, World Health Organization‡ and University of California, Berkeley§*

Estimating intrapartum stillbirths (IPSB) fraction

- The model is constructed very carefully with lots of thoughts going into the model component, specification, and how to properly account for data quality.
- Some similar efforts in modeling key demographic indicators:

Received: 16 October 2020 | Revised: 11 November 2021 | Accepted: 9 December 2021

DOI: 10.1002/sim.9335

RESEARCH ARTICLE

Statistics in Medicine WILEY

Estimating misclassification errors in the reporting of maternal mortality in national civil registration vital statistics systems: A Bayesian hierarchical bivariate random walk model to estimate sensitivity and specificity for multiple countries and years with missing data

Emily Peterson[1] | Doris Chou[2] | Ann-Beth Moller[2] | Alison Gemmill[3] | Lale Say[2] | Leontine Alkema[4]

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1245–1274
DOI: 10.1214/16-AOAS1014
© Institute of Mathematical Statistics, 2017

A BAYESIAN APPROACH TO THE GLOBAL ESTIMATION OF MATERNAL MORTALITY[1]

BY LEONTINE ALKEMA*, SANQIAN ZHANG†, DORIS CHOU‡, ALISON GEMMILL§, ANN-BETH MOLLER‡, DORIS MA FAT‡, LALE SAY‡, COLIN MATHERS‡ AND DANIEL HOGAN‡

University of Massachusetts, Amherst, Harvard University†, World Health Organization‡ and University of California, Berkeley§*

Estimating intrapartum stillbirths (IPSB) fraction

- The model is constructed very carefully with lots of thoughts going into the model component, specification, and how to properly account for data quality.
- Some similar efforts in modeling key demographic indicators:

Received: 16 October 2020 | Revised: 11 November 2021 | Accepted: 9 December 2021

DOI: 10.1002/sim.9335

RESEARCH ARTICLE

Statistics in Medicine

Estimating misclassification errors in the reporting of maternal mortality in national civil registration vital statistics systems: A Bayesian hierarchical bivariate ra walk model to estimate sensitivity and specificity for multiple countries and years with missing data

Emily Peterson[1] | Doris Chou[2] | Ann-Beth Moller[2] | Alison Gemmill[3] | Lale Say[2] | Leontine Alkema[4]

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1245–1274
DOI: 10.1214/16-AOAS1014

Sex differences in mortality among children, adolescents, and young people aged 0–24 years: a systematic assessment of national, regional, and global trends from 1990 to 2021

Fengqing Chao, Bruno Masquelier, Danzhen You, Lucia Hug, Yang Liu, David Sharrow, Håvard Rue, Hernando Ombao, and Leontine Alkema, on behalf of the UN Inter-agency Group for Child Mortality Estimation

Estimating intrapartum stillbirths (IPSB) fraction

- The model is constructed very carefully with lots of thoughts going into the model component, specification, and how to properly account for data quality.
- Some similar efforts in modeling key demographic indicators:

RESEARCH ARTICLE

Changes in the spatial distribution of the under-five mortality rate: Small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa

Zehang Li[1☯], Yuan Hsiao[2,4☯], Jessica Godwin[2☯], Bryan D. Martin[2☯], Jon Wakefield[2,3☯‡], Samuel J. Clark[5,6☯‡*], with support from the United Nations Inter-agency Group for Child Mortality Estimation and its technical advisory group[¶]

ed Statistics
, 1245–1274
DAS1014

ıces in mortality among children, adolescents, people aged 0–24 years: a systematic assessment , regional, and global trends from 1990 to 2021

quelier, Danzhen You, Lucia Hug, Yang Liu, David Sharrow, Håvard Rue, Hernando Ombao, and Leontine Alkema, on ıcy Group for Child Mortality Estimation

Estimating intrapartum stillbirths (IPSB) fraction

- The model is constructed very carefully with lots of thoughts going into the model component, specification, and how to properly account for data quality.
- Some similar efforts in modeling key demographic indicators:

RESEARCH ARTICLE

Changes in the spatial distribution of the under-five mortality rate: Small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa

Zehang Li[1☯], Yuan Hsiao[2,4☯], Jessica Godwin[2☯], Bryan D. Martin[2☯], Jon Wakefield[2,3☯‡], Samuel J. Clark[5,6☯‡*], with support from the United Nations Inter-agency Group for Child Mortality Estimation and its technical advisory group[¶]

Spatial Modeling for Subnational Administrative Level 2 Small-Area Estimation

Yunhan Wu,[1] Zehang Richard Li,[2] Benjamin K. Mayala,[3] Houjie Wang,[1] Peter A. Gao,[1] John Paige,[4] Geir-Arne Fuglstad,[4] Caitlin Moe,[1] Jessica Godwin,[1] Rose E. Donohue,[3] Bradley Janocha,[3] Trevor N. Croft,[3] and Jon Wakefield[1,5]

The DHS Program
ICF
Rockville, Maryland, USA

September 2021

How much information to share?

How much information to share?

- The key challenge to me is how much information to share and what assumptions to make.

How much information to share?

- The key challenge to me is how much information to share and what assumptions to make.

- In our work on subnational child mortality estimation and small area estimation, we usually choose to model data from a single country.

How much information to share?

- The key challenge to me is how much information to share and what assumptions to make.
- In our work on subnational child mortality estimation and small area estimation, we usually choose to model data from a single country.

model. Assuming constant hazards within age bands, we assume the number of deaths occurring within age band $a[m]$, in cluster c, time t, and survey k follow the beta-binomial distribution,

$$Y_{a[m],k,c,t} \mid p_{a[m],k,c,t} \sim \text{BetaBinomial}\left(n_{a[m],k,c,t} , p_{m,k,c,t} , d \right), \tag{7}$$

where $p_{m,k,c,t}$ is the monthly hazard at m-th month of age, in cluster c, time t, and survey k and d is the overdispersion parameter. The latent logistic model we use is,

$$p_{m,k,c,t} = \text{expit}(\alpha_{m,c,k,t} + \epsilon_t + b_k), \tag{8}$$

$$\alpha_{m,k,c,t} = \beta_{a^\star[m],r[k],t} I(s_c \in \text{ rural }) + \gamma_{a^\star[m],r[k],t} I(s_c \in \text{ urban }) + S_{i[s_c]} + e_{i[s_c]} + \delta_{i[s_c],t} + \text{BIAS}_{k,t}. \tag{9}$$

How much information to share?

- The key challenge to me is how much information to share and what assumptions to make.
- In our work on subnational child mortality estimation and small area estimation, we usually choose to model data from a single country.

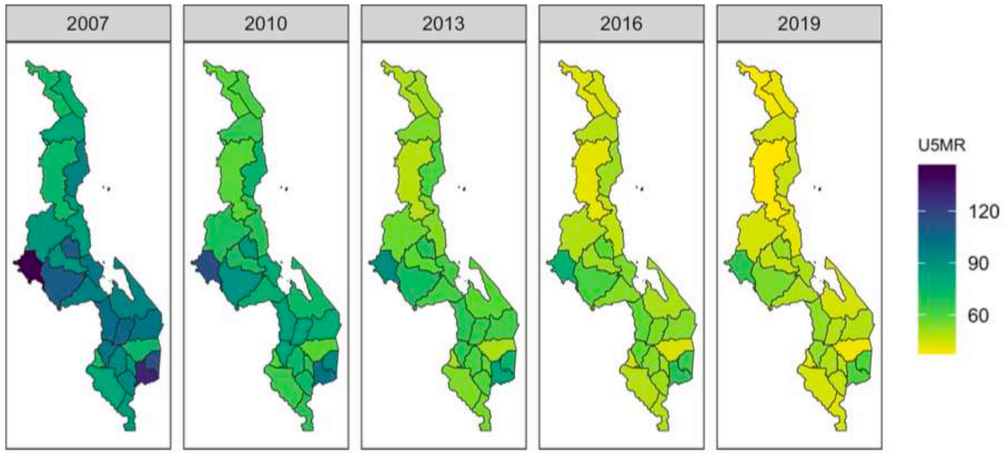
model. Assuming constant hazards within age bands, we assume the number of deaths occurring within age band $a[m]$, in cluster c, time t, and survey k follow the beta-binomial distribution,

$$Y_{a[m],k,c,t} \mid p_{a[m],k,c,t} \sim \text{BetaBinomial}\left(n_{a[m],k,c,t}, p_{m,k,c,t}, d\right), \tag{7}$$

where $p_{m,k,c,t}$ is the monthly hazard at m-th month of age, in cluster c, time t, and survey k and d is the overdispersion parameter. The latent logistic model we use is,

$$p_{m,k,c,t} = \text{expit}(\alpha_{m,c,k,t} + \epsilon_t + b_k), \tag{8}$$

$$\alpha_{m,k,c,t} = \beta_{a^\star[m],r[k],t} I(s_c \in \text{rural}) + \gamma_{a^\star[m],r[k],t} I(s_c \in \text{urban}) + S_{i[s_c]} + e_{i[s_c]} + \delta_{i[s_c],t} + \text{BIAS}_{k,t}. \tag{9}$$



How much information to share?

- The key challenge to me is how much information to share and what assumptions to make.
- In our work on subnational child mortality estimation and small area estimation, we usually choose to model data from a single country.

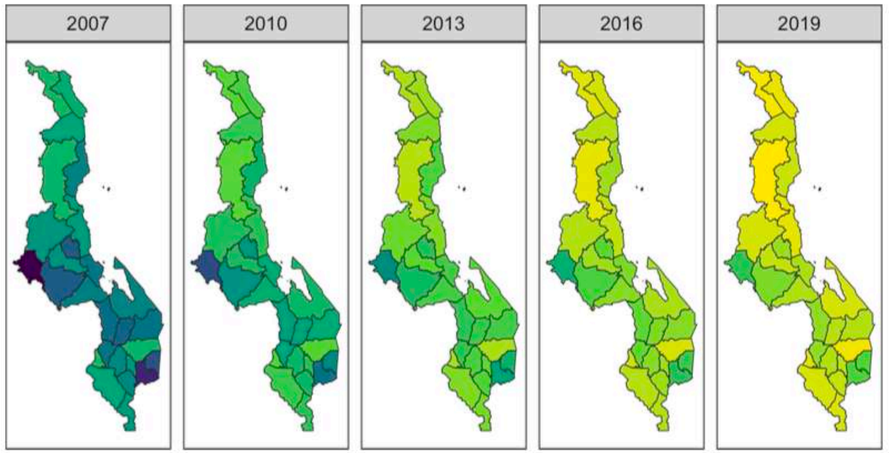
model. Assuming constant hazards within age bands, we assume the number of deaths occurring within age band $a[m]$, in cluster c, time t, and survey k follow the beta-binomial distribution,

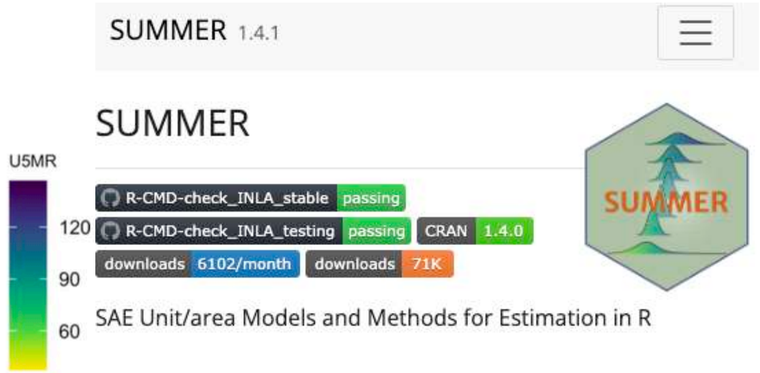
$$Y_{a[m],k,c,t} \mid p_{a[m],k,c,t} \sim \text{BetaBinomial}\left(n_{a[m],k,c,t}, p_{m,k,c,t}, d\right), \tag{7}$$

where $p_{m,k,c,t}$ is the monthly hazard at m-th month of age, in cluster c, time t, and survey k and d is the overdispersion parameter. The latent logistic model we use is,

$$p_{m,k,c,t} = \text{expit}(\alpha_{m,c,k,t} + \epsilon_t + b_k), \tag{8}$$

$$\alpha_{m,k,c,t} = \beta_{a^\star[m],r[k],t} I(s_c \in \text{rural}) + \gamma_{a^\star[m],r[k],t} I(s_c \in \text{urban}) + S_{i[s_c]} + e_{i[s_c]} + \delta_{i[s_c],t} + \text{BIAS}_{k,t}. \tag{9}$$





How much information to share?

- The key challenge to me is how much information to share and what assumptions to make.
- In our work on subnational child mortality estimation and small area estimation, we usually choose to model data from a single country.

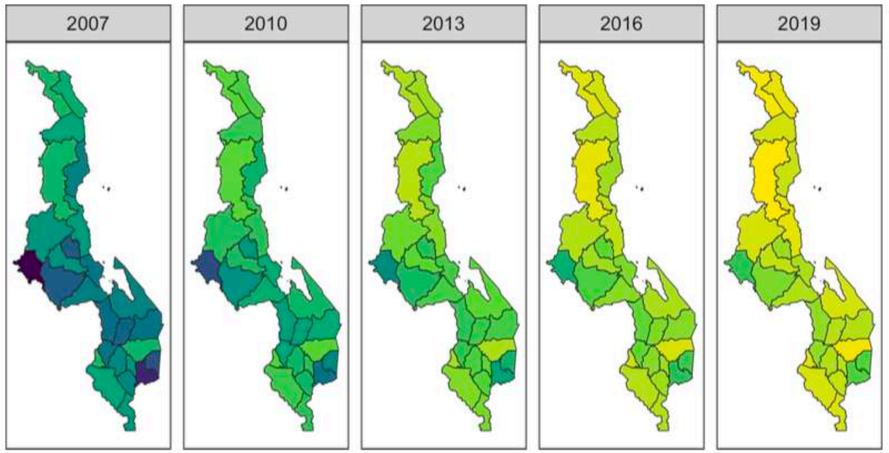
model. Assuming constant hazards within age bands, we assume the number of deaths occurring within age band $a[m]$, in cluster c, time t, and survey k follow the beta-binomial distribution,

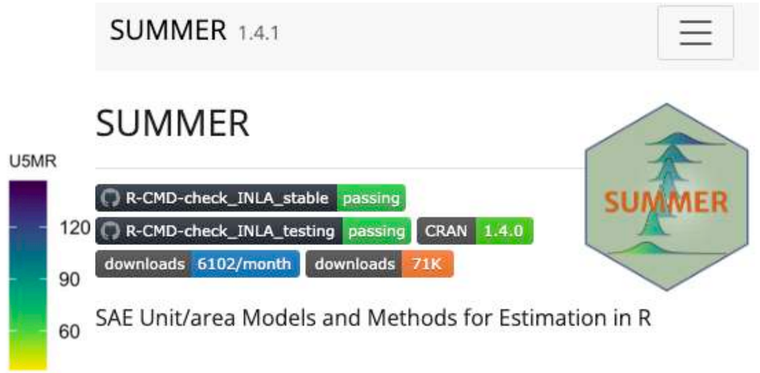
$$Y_{a[m],k,c,t} \mid p_{a[m],k,c,t} \sim \text{BetaBinomial}\left(n_{a[m],k,c,t}, p_{m,k,c,t}, d\right), \tag{7}$$

where $p_{m,k,c,t}$ is the monthly hazard at m-th month of age, in cluster c, time t, and survey k and d is the overdispersion parameter. The latent logistic model we use is,

$$p_{m,k,c,t} = \text{expit}(\alpha_{m,c,k,t} + \epsilon_t + b_k), \tag{8}$$

$$\alpha_{m,k,c,t} = \beta_{a^\star[m],r[k],t} I(s_c \in \text{rural}) + \gamma_{a^\star[m],r[k],t} I(s_c \in \text{urban}) + S_{i[s_c]} + e_{i[s_c]} + \delta_{i[s_c],t} + \text{BIAS}_{k,t}. \tag{9}$$





https://richardli.github.io/SUMMER/

How much information to share?

- The key challenge to me is how much information to share and what assumptions to make.
- In our work on subnational child mortality estimation and small area estimation, we usually choose to model data from a single country.
- For summary data, global model is probably the only choice? How do we ensure the model extrapolates in a sensible way across countries?

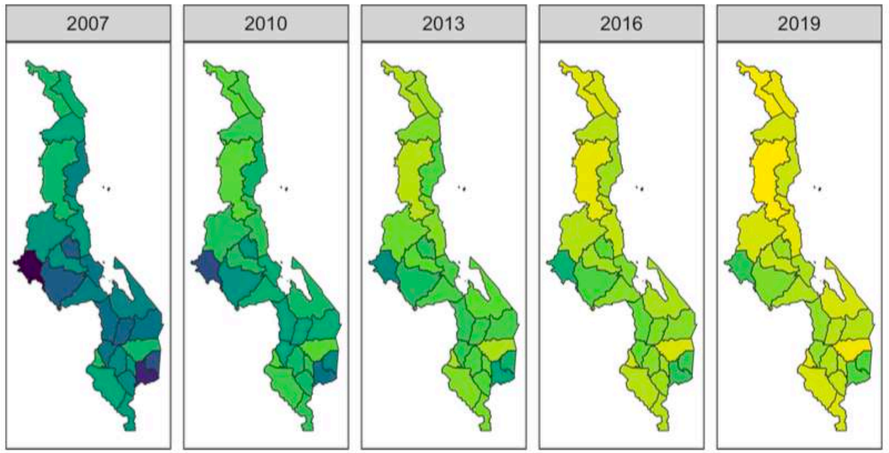
model. Assuming constant hazards within age bands, we assume the number of deaths occurring within age band $a[m]$, in cluster c, time t, and survey k follow the beta-binomial distribution,

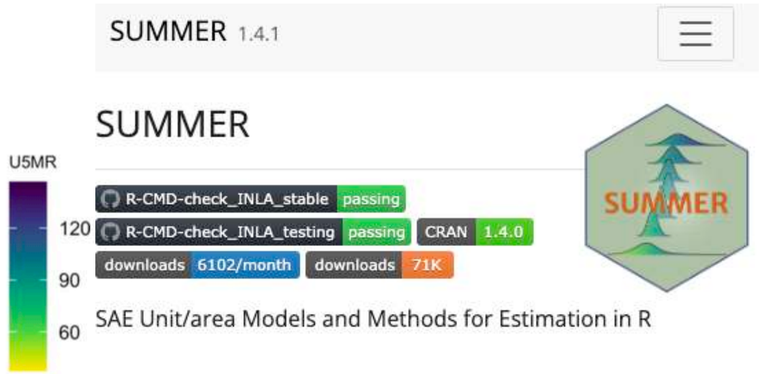
$$Y_{a[m],k,c,t} \mid p_{a[m],k,c,t} \sim \text{BetaBinomial}\left(n_{a[m],k,c,t}, p_{m,k,c,t}, d\right), \tag{7}$$

where $p_{m,k,c,t}$ is the monthly hazard at m-th month of age, in cluster c, time t, and survey k and d is the overdispersion parameter. The latent logistic model we use is,

$$p_{m,k,c,t} = \text{expit}(\alpha_{m,c,k,t} + \epsilon_t + b_k), \tag{8}$$

$$\alpha_{m,k,c,t} = \beta_{a^*[m],r[k],t} I(s_c \in \text{rural}) + \gamma_{a^*[m],r[k],t} I(s_c \in \text{urban}) + S_{i[s_c]} + e_{i[s_c]} + \delta_{i[s_c],t} + \text{BIAS}_{k,t}. \tag{9}$$





https://richardli.github.io/SUMMER/

How much information to share?

- The key challenge to me is how much information to share and what assumptions to make.
- In our work on subnational child mortality estimation and small area estimation, we usually choose to model data from a single country.
- For summary data, global model is probably the only choice? How do we ensure the model extrapolates in a sensible way across countries?
- How do we assess the amount of information shared and avoid "over-smoothing"?

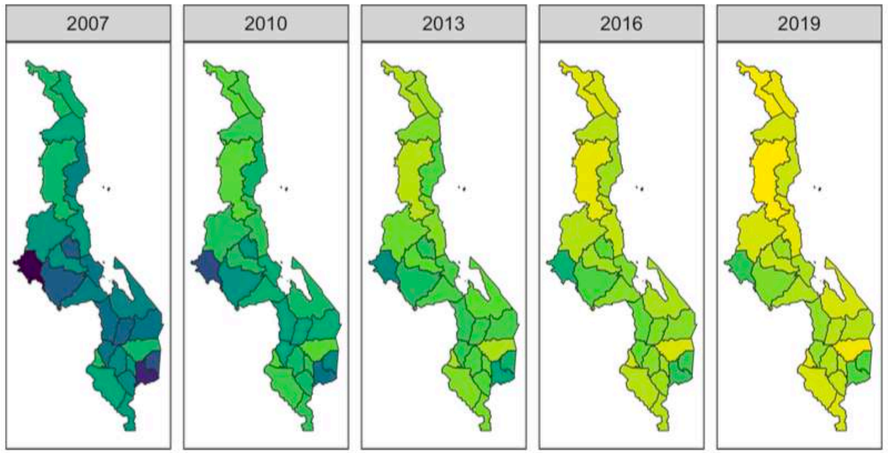
model. Assuming constant hazards within age bands, we assume the number of deaths occurring within age band $a[m]$, in cluster c, time t, and survey k follow the beta-binomial distribution,

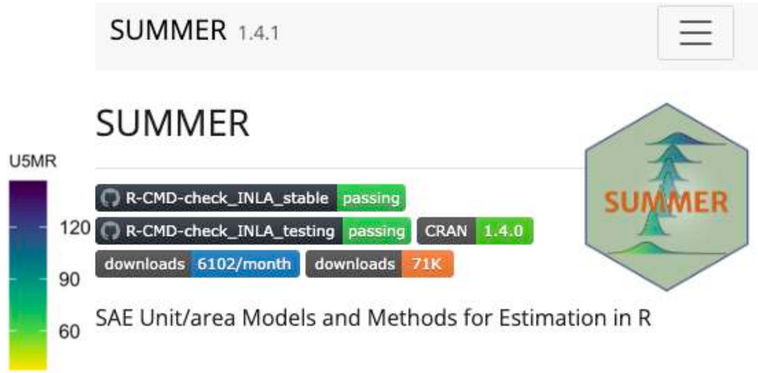
$$Y_{a[m],k,c,t} \mid p_{a[m],k,c,t} \sim \text{BetaBinomial}\left(n_{a[m],k,c,t}, p_{m,k,c,t}, d\right), \tag{7}$$

where $p_{m,k,c,t}$ is the monthly hazard at m-th month of age, in cluster c, time t, and survey k and d is the overdispersion parameter. The latent logistic model we use is,

$$p_{m,k,c,t} = \text{expit}(\alpha_{m,c,k,t} + \epsilon_t + b_k), \tag{8}$$

$$\alpha_{m,k,c,t} = \beta_{a^*[m],r[k],t} I(s_c \in \text{rural}) + \gamma_{a^*[m],r[k],t} I(s_c \in \text{urban}) + S_{i[s_c]} + e_{i[s_c]} + \delta_{i[s_c],t} + \text{BIAS}_{k,t}. \tag{9}$$





https://richardli.github.io/SUMMER/

The importance of having estimates

The importance of having estimates

- It is very important to have estimates for key demographic indicators.

The importance of having estimates

- It is very important to have estimates for key demographic indicators.

- But on the other hand, policy-making based on estimates with high uncertainty is problematic.

The importance of having estimates

- It is very important to have estimates for key demographic indicators.

- But on the other hand, policy-making based on estimates with high uncertainty is problematic.

- How do we convey uncertainty in the model output to policy makers?

The importance of having estimates

- It is very important to have estimates for key demographic indicators.
- But on the other hand, policy-making based on estimates with high uncertainty is problematic.
- How do we convey uncertainty in the model output to policy makers?

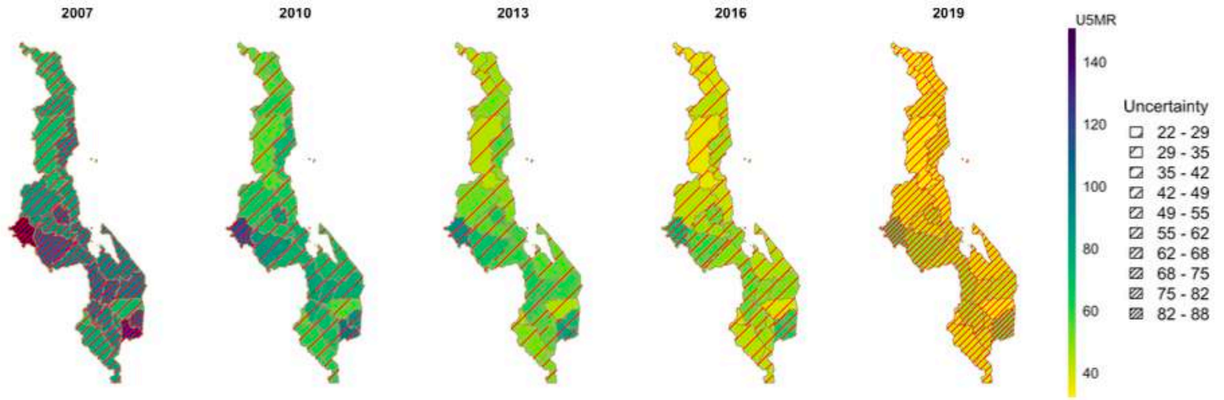


Figure 6: Subnational estimates of U5MR using the 2015–2016 DHS in Malawi over selected years, with hatching lines indicating the width of the 95% credible intervals of the estimates. Denser hatching correspond to higher uncertainty. Estimates for 2019 in the last column are from the model projection and thus have higher uncertainty.

The importance of having estimates

- It is very important to have estimates for key demographic indicators.
- But on the other hand, policy-making based on estimates with high uncertainty is problematic.
- How do we convey uncertainty in the model output to policy makers?

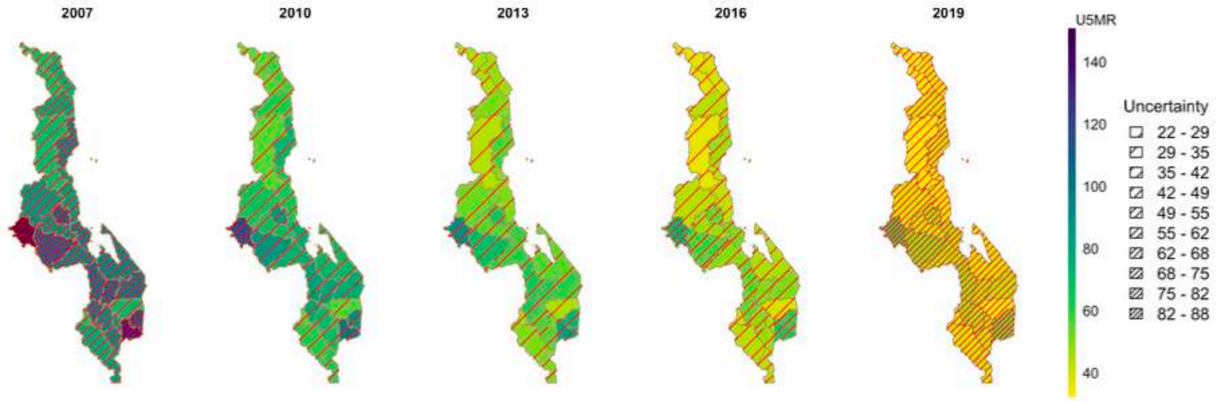
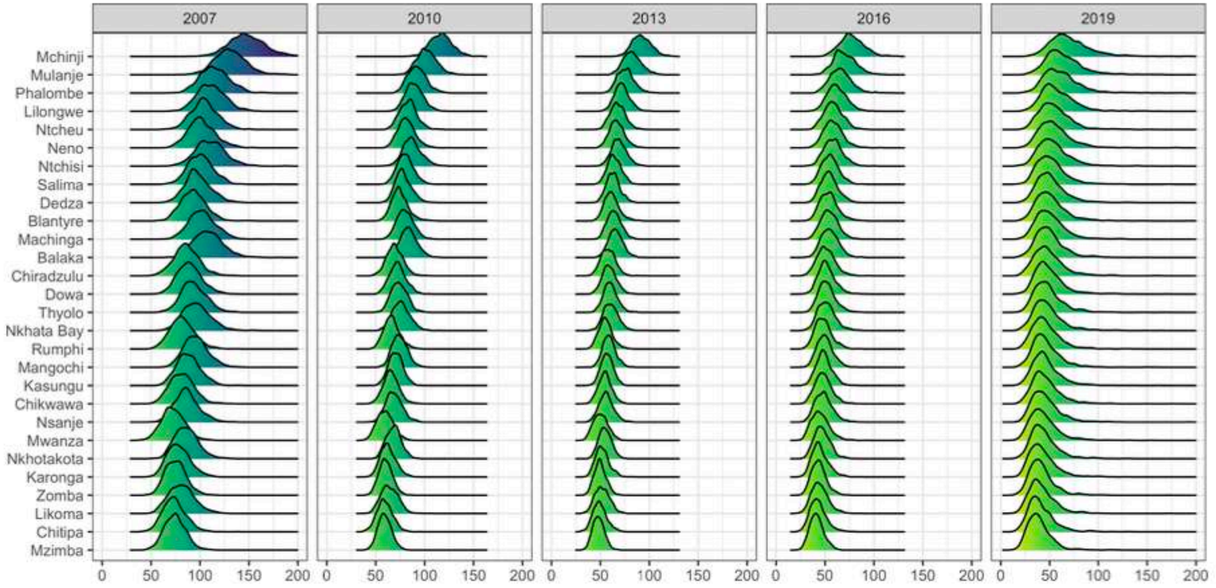


Figure 6: Subnational estimates of U5MR using the 2015–2016 DHS in Malawi over selected years, with hatching lines indicating the width of the 95% credible intervals of the estimates. Denser hatching correspond to higher uncertainty. Estimates for 2019 in the last column are from the model projection



The need for aggregation

The need for aggregation

- Subpopulation-level model allows researchers to combine data from multiple sources, with different resolutions, and allows covariate modeling at the relevant scale.

The need for aggregation

- Subpopulation-level model allows researchers to combine data from multiple sources, with different resolutions, and allows covariate modeling at the relevant scale.

- Aggregating subpopulation estimates to the useful scale can be non-trivial.

The need for aggregation

- Subpopulation-level model allows researchers to combine data from multiple sources, with different resolutions, and allows covariate modeling at the relevant scale.

- Aggregating subpopulation estimates to the useful scale can be non-trivial.

surveyPrev: Mapping the Prevalence of Binary Indicators using Survey Data in Small Areas

Provides a pipeline to perform small area estimation and prevalence mapping of binary indicators using health and demographic survey data, described in Fuglstad et al. (2022) <doi:10.48550/arXiv.2110.09576> and Wakefield et al. (2020) <doi:10.1111/insr.12400>.

Version:	1.0.0
Depends:	R (≥ 3.5)
Imports:	survey, stats, ggplot2, rdhs, SUMMER, dplyr, labelled, sjlabelled, naniar, raster, sp, spdep, stringr, tidyverse, data.table, sf, matrixStats
Suggests:	INLA, knitr, rmarkdown, R.rsp, kableExtra, geodata, patchwork, tidyr
Published:	2024-04-10
Author:	Qianyu Dong [cre, aut], Zehang R Li [aut], Yunhan Wu [aut], Andrea Boskovic [aut], Jon Wakefield [aut]

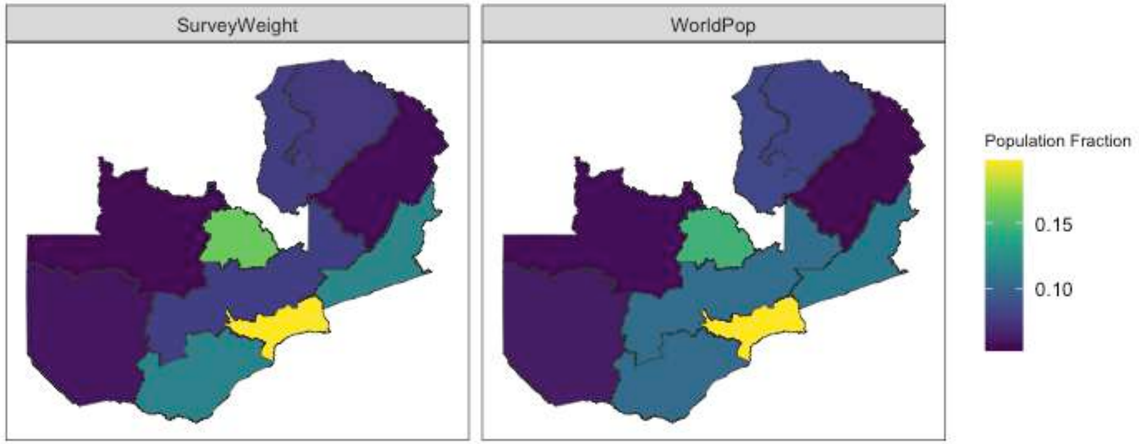
The need for aggregation

- Subpopulation-level model allows researchers to combine data from multiple sources, with different resolutions, and allows covariate modeling at the relevant scale.
- Aggregating subpopulation estimates to the useful scale can be non-trivial.

surveyPrev: Mapping the Prevalence of Binary Indicators using Survey Data in Small Areas

Provides a pipeline to perform small area estimation and prevalence mapping of binary indicators using health and demographic survey data, described in Fuglstad et al. (2022) <doi:10.48550/arXiv.2110.09576> and Wakefield et al. (2020) <doi:10.1111/insr.12400>.

Version:	1.0.0
Depends:	R (≥ 3.5)
Imports:	survey, stats, ggplot2, rdhs, SUMMER, dplyr, labelled, sjlabelled, naniar, raster, sp, spdep, stringr, tidyverse, data.table, sf, matrixStats
Suggests:	INLA, knitr, rmarkdown, R.rsp, kableExtra, geodata, patchwork, tidyr
Published:	2024-04-10
Author:	Qianyu Dong [cre, aut], Zehang R Li [aut], Yunhan Wu [aut], Andrea Boskovic [aut], Jon Wakefield [aut]



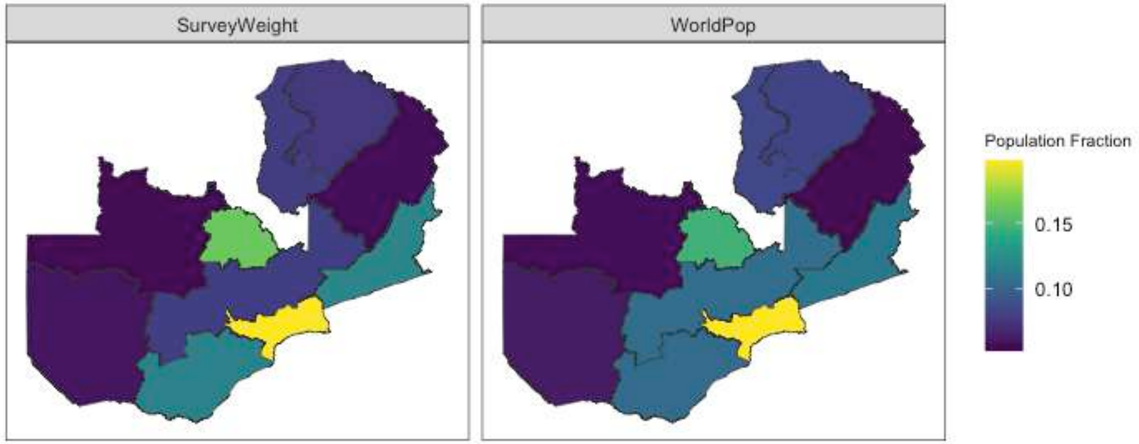
The need for aggregation

- Subpopulation-level model allows researchers to combine data from multiple sources, with different resolutions, and allows covariate modeling at the relevant scale.
- Aggregating subpopulation estimates to the useful scale can be non-trivial.

surveyPrev: Mapping the Prevalence of Binary Indicators using Survey Data in Small Areas

Provides a pipeline to perform small area estimation and prevalence mapping of binary indicators using health and demographic survey data, described in Fuglstad et al. (2022) <doi:10.48550/arXiv.2110.09576> and Wakefield et al. (2020) <doi:10.1111/insr.12400>.

Version:	1.0.0
Depends:	R (≥ 3.5)
Imports:	survey, stats, ggplot2, rdhs, SUMMER, dplyr, labelled, sjlabelled, naniar, raster, sp, spdep, stringr, tidyverse, data.table, sf, matrixStats
Suggests:	INLA, knitr, rmarkdown, R.rsp, kableExtra, geodata, patchwork, tidyr
Published:	2024-04-10
Author:	Qianyu Dong [cre, aut], Zehang R Li [aut], Yunhan Wu [aut], Andrea Boskovic [aut], Jon Wakefield [aut]



https://cran.r-project.org/web/packages/surveyPrev/index.html

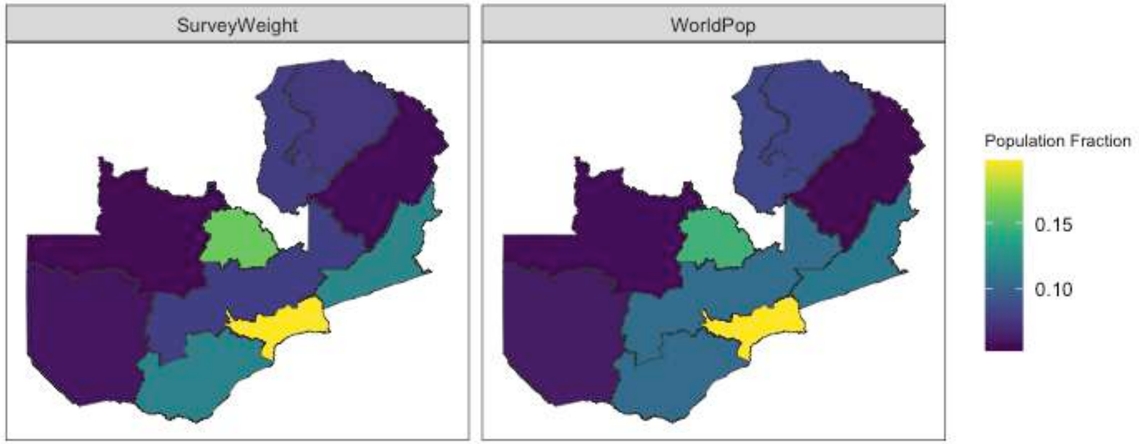
The need for aggregation

- Subpopulation-level model allows researchers to combine data from multiple sources, with different resolutions, and allows covariate modeling at the relevant scale.
- Aggregating subpopulation estimates to the useful scale can be non-trivial.
- How do we assess the effect from aggregation weights?

surveyPrev: Mapping the Prevalence of Binary Indicators using Survey Data in Small Areas

Provides a pipeline to perform small area estimation and prevalence mapping of binary indicators using health and demographic survey data, described in Fuglstad et al. (2022) <doi:10.48550/arXiv.2110.09576> and Wakefield et al. (2020) <doi:10.1111/insr.12400>.

Version:	1.0.0
Depends:	R (≥ 3.5)
Imports:	survey, stats, ggplot2, rdhs, SUMMER, dplyr, labelled, sjlabelled, naniar, raster, sp, spdep, stringr, tidyverse, data.table, sf, matrixStats
Suggests:	INLA, knitr, rmarkdown, R.rsp, kableExtra, geodata, patchwork, tidyr
Published:	2024-04-10
Author:	Qianyu Dong [cre, aut], Zehang R Li [aut], Yunhan Wu [aut], Andrea Boskovic [aut], Jon Wakefield [aut]



https://cran.r-project.org/web/packages/surveyPrev/index.html

Verbal Autopsy

Verbal Autopsy

- VA is usually the only feasible method to collect information on cause of death where traditional death certification or autopsy are not possible.

Verbal Autopsy

- VA is usually the only feasible method to collect information on cause of death where traditional death certification or autopsy are not possible.
- Zhenke's paper deals with the important problem of distribution shift across datasets.

Verbal Autopsy

- VA is usually the only feasible method to collect information on cause of death where traditional death certification or autopsy are not possible.
- Zhenke's paper deals with the important problem of distribution shift across datasets.
- This is also a general problem for any predictive modeling in demographic and health research.

Tree-structured domain adaptation

Tree-structured domain adaptation

- The key idea here is that that there are multiple types of symptom profiles for any given cause of death.

Tree-structured domain adaptation

- The key idea here is that that there are multiple types of symptom profiles for any given cause of death.

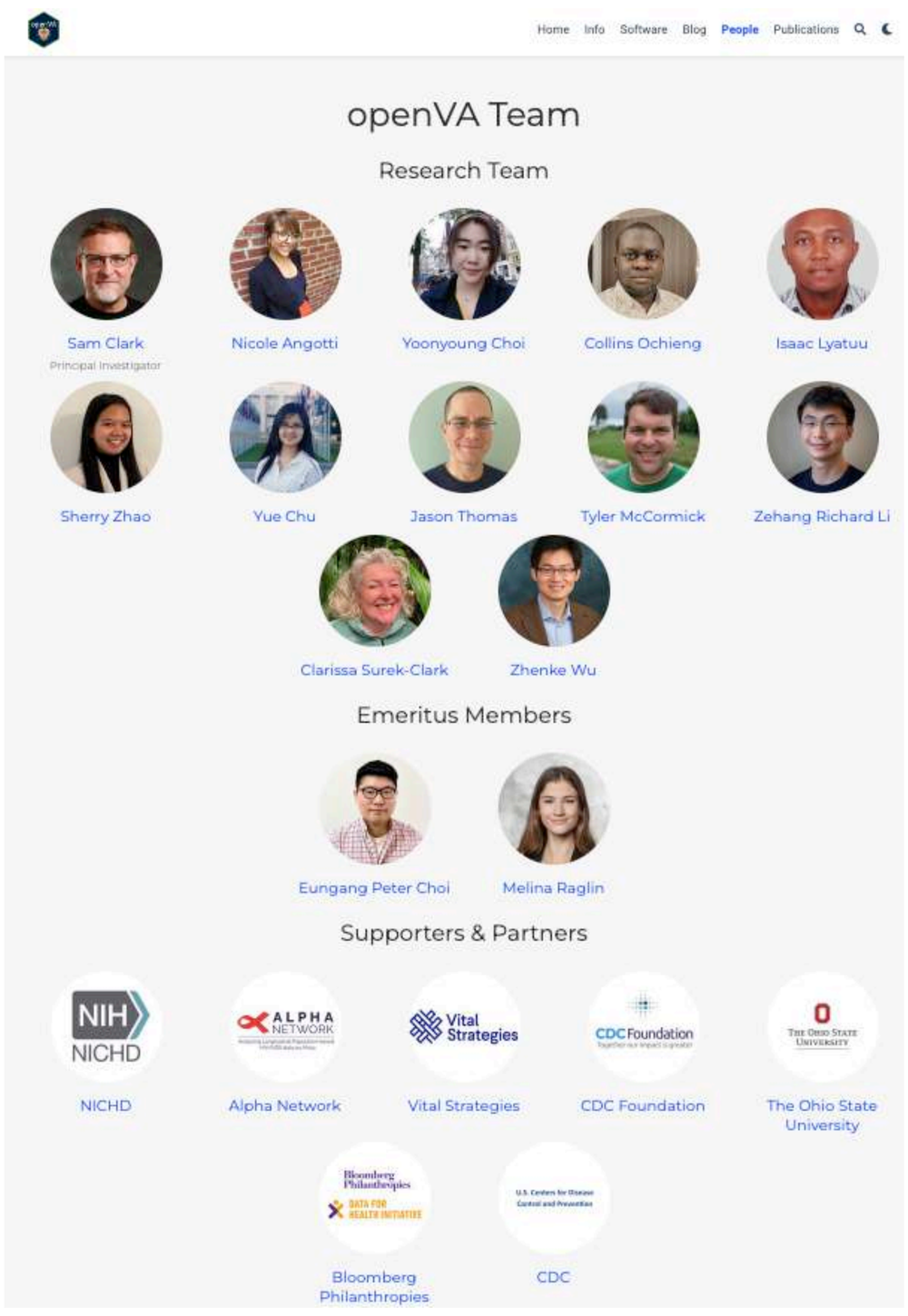
- The observed symptom distribution given each cause of death is a mixture of these latent profiles, and thus can be heterogeneous across populations.

Tree-structured domain adaptation

- The key idea here is that that there are multiple types of symptom profiles for any given cause of death.

- The observed symptom distribution given each cause of death is a mixture of these latent profiles, and thus can be heterogeneous across populations.

- The mixing weights of these latent profiles are more likely to be similar if the populations are “close” to each other.



The openVA team develops and maintains various tools, algorithms, and software related to Verbal Autopsy. Development is ongoing, and things are changing rapidly. This site provides information useful for installing and running openVA as well as recent publications and updates on what the team has done. All of our software itself is contained in 'packages' written for the free, open-source statistical programming environment R and Python. The packages are available through the Comprehensive R Archive Network (CRAN) and the Python Package Index (PyPi).

https://openva.net

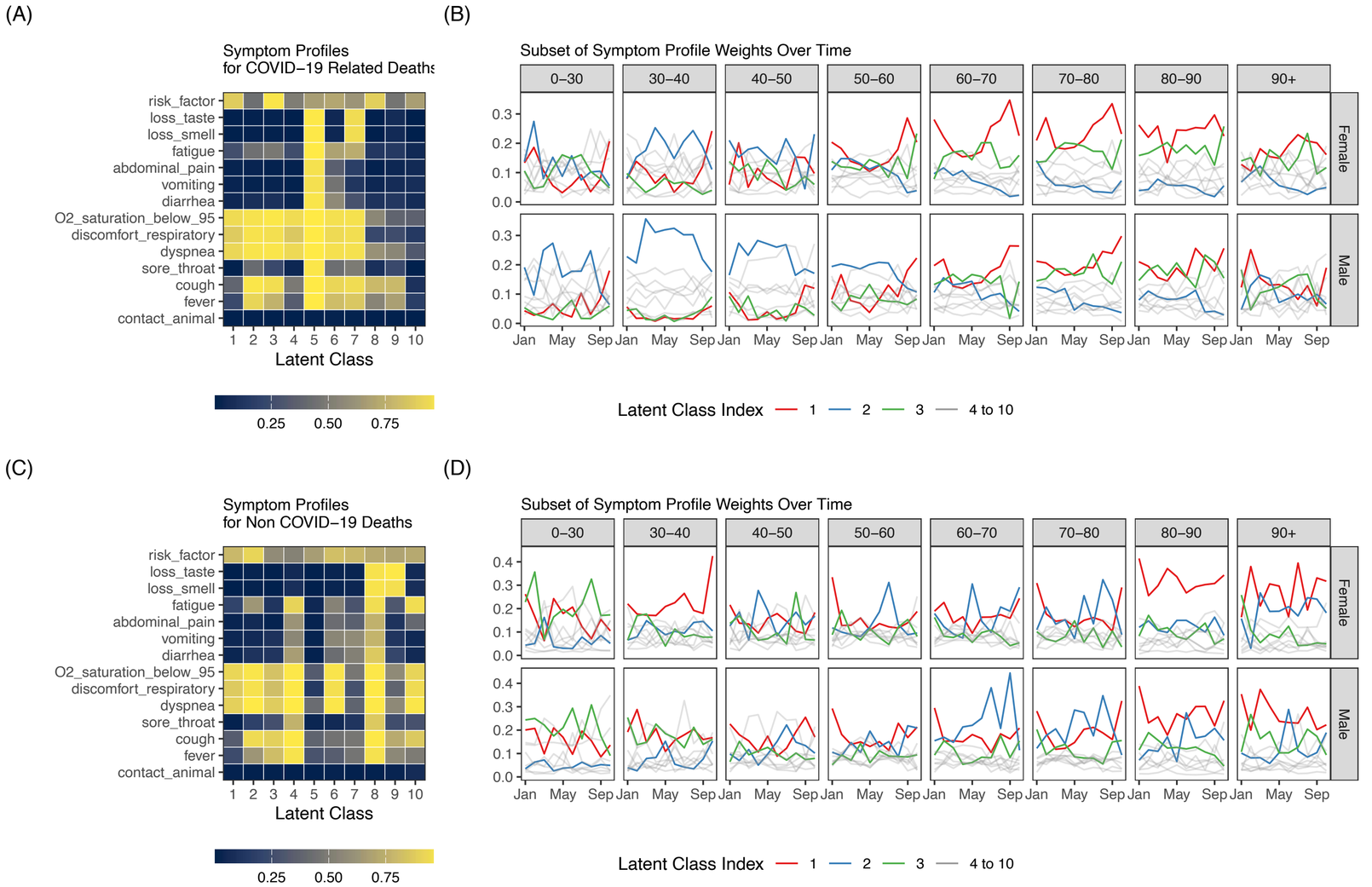
How much information to share

How much information to share

- In related work, we have developed methods for domain adaptation across subpopulation defined by age, sex, time, etc., and use structured prior to smooth the estimates.

How much information to share

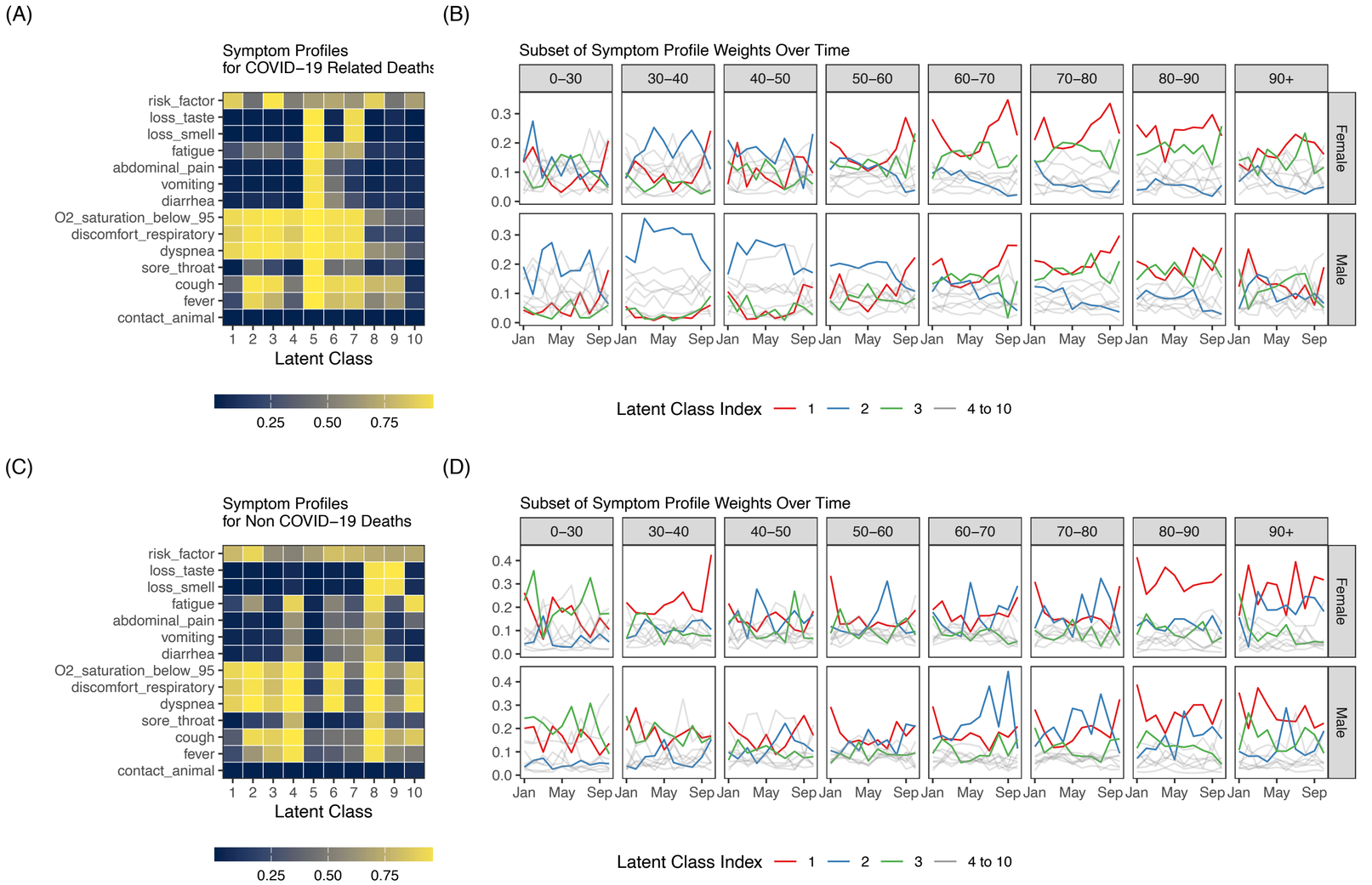
- In related work, we have developed methods for domain adaptation across subpopulation defined by age, sex, time, etc., and use structured prior to smooth the estimates.



How much information to share

- In related work, we have developed methods for domain adaptation across subpopulation defined by age, sex, time, etc., and use structured prior to smooth the estimates.

- What if one or several domains significantly deviate from the structure we assume? How to prevent negative effect from joint modeling?



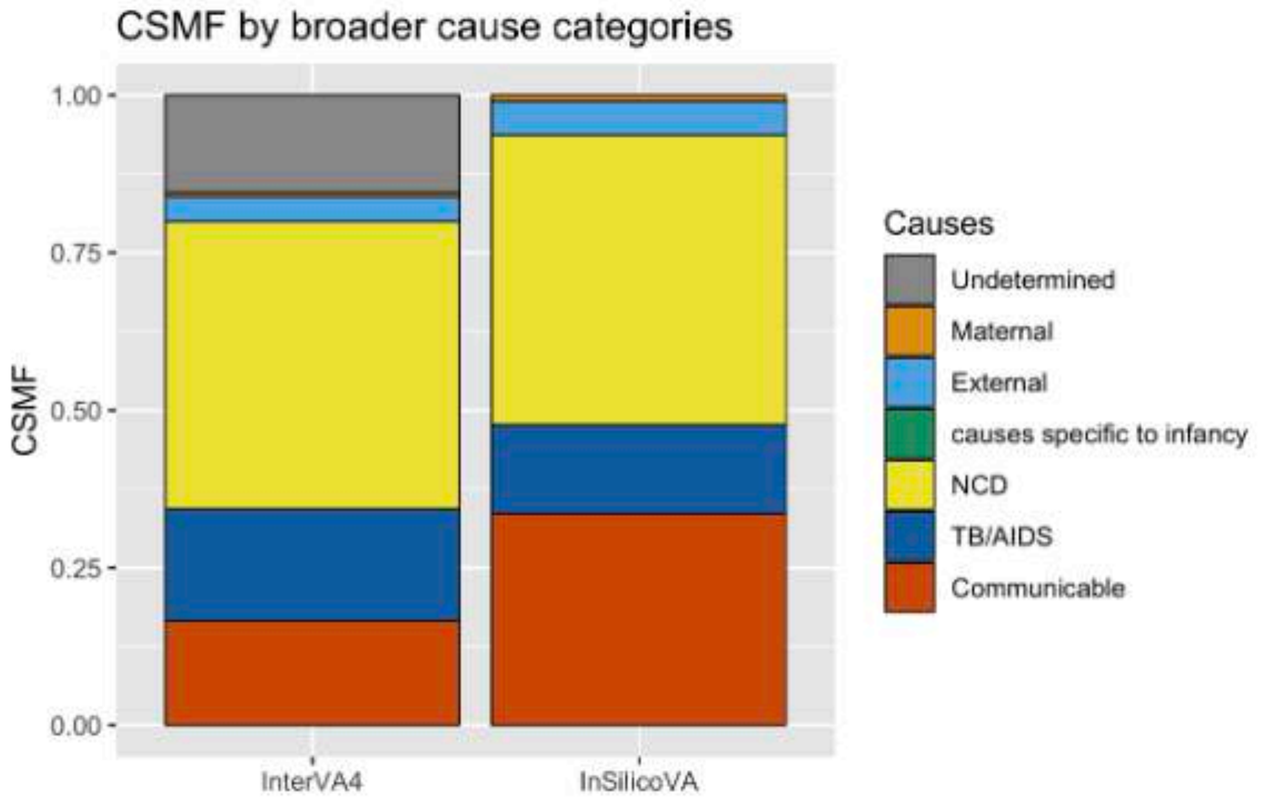
The importance to have estimates

The importance to have estimates

- Early work on VA usually have arbitrary thresholds to report “undetermined” as a cause assignment.

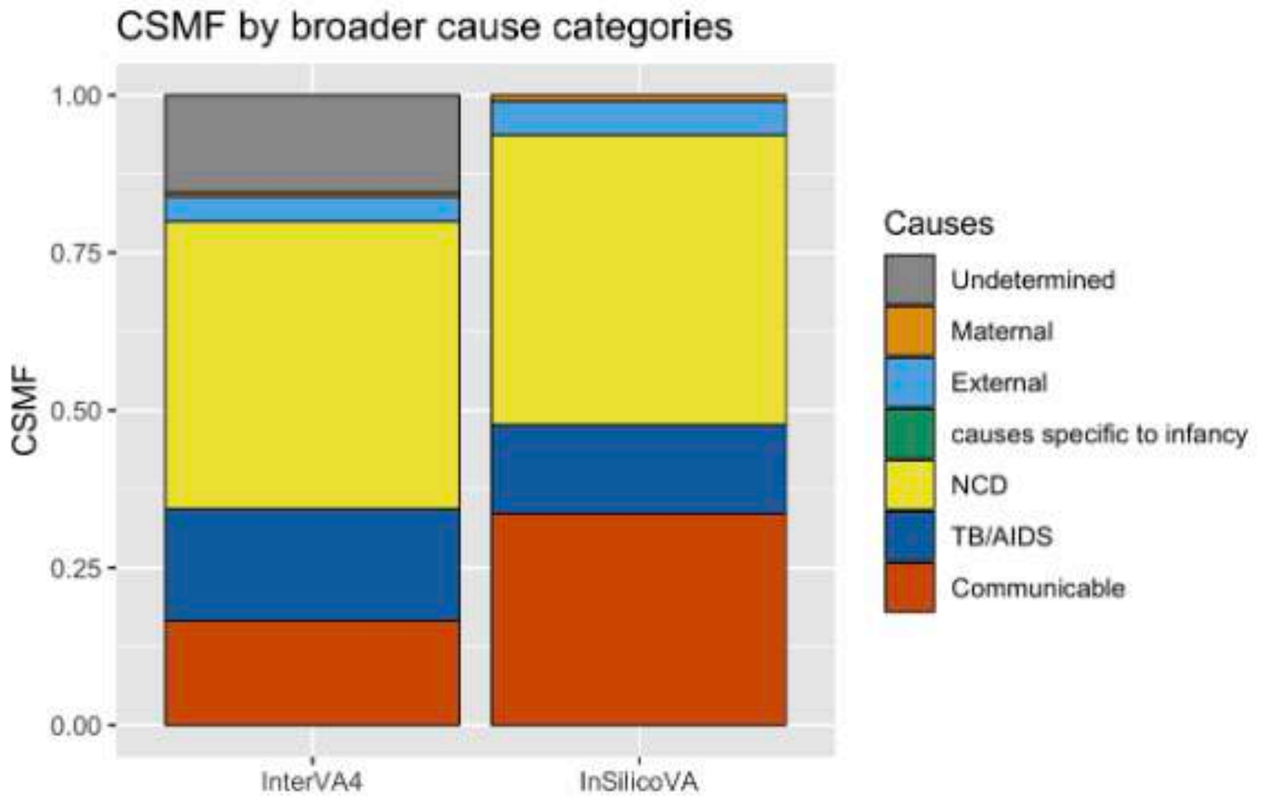
The importance to have estimates

- Early work on VA usually have arbitrary thresholds to report “undetermined” as a cause assignment.



The importance to have estimates

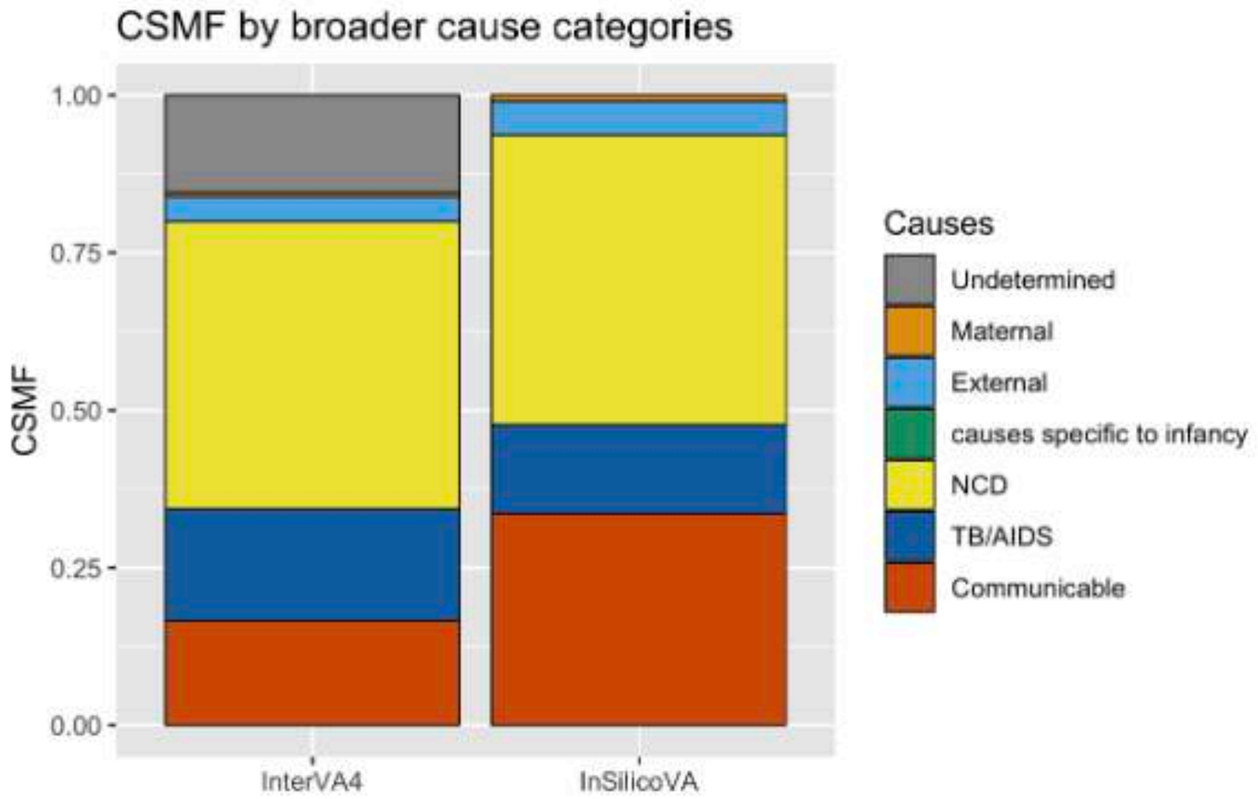
- Early work on VA usually have arbitrary thresholds to report “undetermined” as a cause assignment.



https://cran.r-project.org/web/packages/openVA/index.html

The importance to have estimates

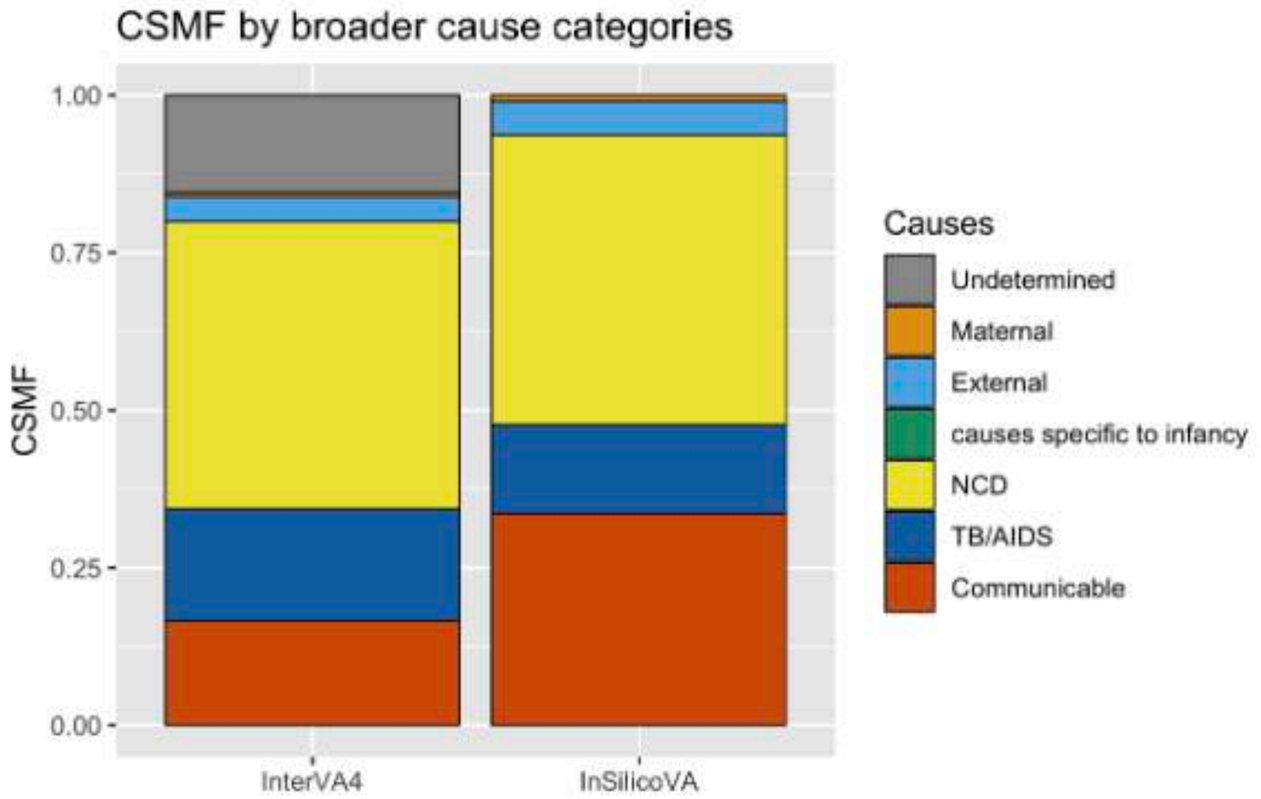
- Early work on VA usually have arbitrary thresholds to report “undetermined” as a cause assignment.
- What if there are “garbage” profiles that correspond to no specific causes?



https://cran.r-project.org/web/packages/openVA/index.html

The importance to have estimates

- Early work on VA usually have arbitrary thresholds to report “undetermined” as a cause assignment.
- What if there are “garbage” profiles that correspond to no specific causes?
- Can we identify them from the latent class model?



https://cran.r-project.org/web/packages/openVA/index.html

The need for aggregation

- How do we combine VAs that are not from a probabilistic sample, with probabilistic surveys or medically certified deaths?

Thank you!

Journal of the Royal Statistical Society Series C: Applied Statistics, 2024, **00**, 1–19
https://doi.org/10.1093/jrsssc/qlae017

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Original Article

Estimating the timing of stillbirths in countries worldwide using a Bayesian hierarchical penalized splines regression model

Michael Y.C. Chong[1] **and Monica Alexander**[1,2]

[1]Department of Statistical Sciences, University of Toronto, Toronto, Canada
[2]Department of Sociology, University of Toronto, Toronto, Canada

Biostatistics, 2024, **00**, 1–21
https://doi.org/10.1093/biostatistics/kxae005
Article

OXFORD

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, United States
[2]Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI 48109, United States
[3]Department of Statistics, University of California, Santa Cruz, CA 95064, United States

*Corresponding author: Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, United States.
Email: zhenkewu@umich.edu

Estimating the timing of stillbirths in all countries

Monica Alexander, Statistical Sciences and Sociology, University of Toronto
IAOS-ISI WSC, 15 May 2024

IGME

unicef
for every child

UNIVERSITY OF TORONTO

Background and Motivation

- Estimated 2 million stillbirths globally
- Reducing stillbirths is an important part of the UN Sustainable Development Goals agenda
- Specific aims to reduce stillbirth rate and end preventable stillbirths

Background and Motivation

- Stillbirths can either occur before or after the onset of labor (antepartum or intrapartum
- Stillbirths that occur intrapartum are largely preventable with adequate access to medical resources and healthcare

***Goal of this project**: estimate the proportion of stillbirths that are intrapartum (IPSB) for all 195 UN-member countries over the period 2000-2021*

Background and Motivation

- There are a number of data quality and availability issues that make estimating IPSB challenging, particularly in low- and middle-income countries
- We use a Bayesian hierarchical penalized splines regression model with a post-estimation weighting step to account for many of these issues
- Just published in JRSS C: https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae017/7636258
- Estimates published in UN report on stillbirths: https://childmortality.org/
- Joint work with Michael Chong (Statistics, UofT), with support and input from members of the UN Interagency Group for Child Mortality Estimation (UN IGME) and UNICEF

Data (or lack thereof)

Characteristics of data on stillbirth timing

Data collection system	• Civil Registration and Vital Statistics (CRVS) system • Health and Medical Information System (HMIS) • Single health facility • Population-based study
Classification method	• Fetal heartbeat • Appearance of skin
Stillbirth definition	• ‘Late’ (official definition): >28 weeks gestation or >1000g • ‘Early’: >22 weeks gestation or >500g

Data availability

- At least one data point for 92 countries
- Big differences in data availability by region
- Big differences in data type by region

Data availability by region

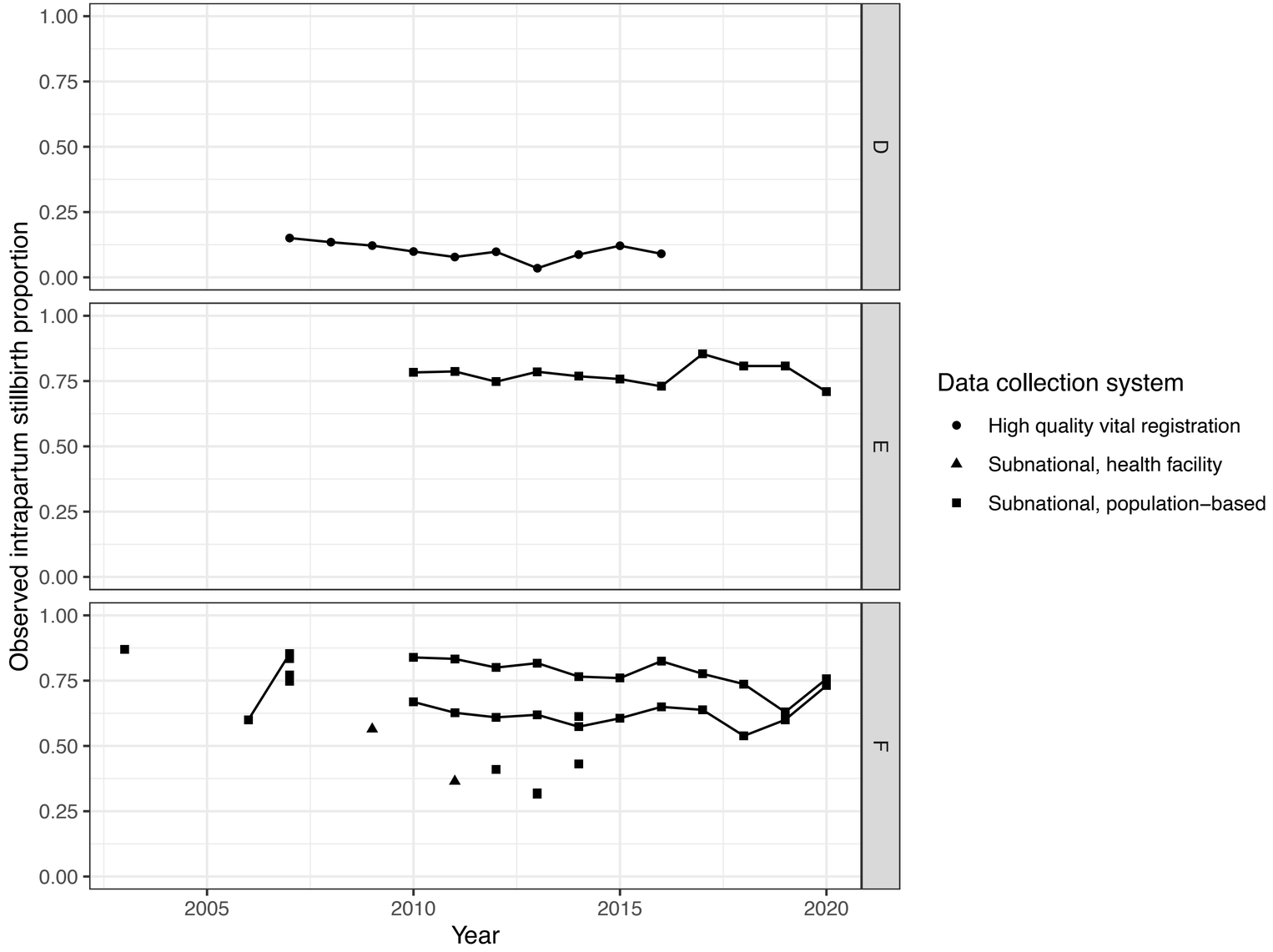
SDG region	Observations	Countries	Country-years
Central and Southern Asia	163	7	65
Eastern and South-Eastern Asia	57	8	53
Latin America and the Caribbean	272	13	171
North America, Europe, Australia and New Zealand	460	30	368
Northern Africa and Western Asia	43	8	42
Oceania (exc. Australia and New Zealand)	1	1	1
Sub-Saharan Africa	280	25	158

Proportion of observations by data collection system

SDG region	CRVS	Health facility	Subnat pop-based	HMIS
Central and Southern Asia	0.067	0.110	0.810	0.012
Eastern and South-Eastern Asia	0.684	0.281	0.035	0.000
Latin America and the Caribbean	0.794	0.044	0.162	0.000
North America, Europe, Australia and New Zealand	0.980	0.015	0.004	0.000
Northern Africa and Western Asia	0.465	0.047	0.093	0.395
Oceania (exc. Australia and New Zealand)	0.000	1.000	0.000	0.000
Sub-Saharan Africa	0.007	0.139	0.375	0.479

Illustrative countries

- Often only have one or two points (no idea of trends)
- Even when countries have data, large variation in type, level and trends



Modeling approach

Modeling goals

- Allow for different levels of measurement error based on data system
- Obtain estimates over time in the absence of temporal data
- Data-driven trends in presence of reliable temporal data
- Account for different stillbirth definitions
- Adjust for under coverage

Model set up

- Consider data on stillbirths by timing as available for a specific ‘place’
- For observations $i = 1, \ldots, N$ let y_i and z_i denote the number of observed intrapartum and antepartum stillbirths respectively. Then

$$y_i \,|\, \phi_i \sim \text{Binomial}(y_i + z_i, \phi_i)$$

- The ϕ_i represents the proportion of intrapartum stillbirths, to be estimated.

Data model

The proportion ϕ_i is modeled

$$\text{logit}(\phi_i) = \mu_i + \varepsilon_i$$

with $\varepsilon_i \sim \text{Normal}(0, \sigma^2_{\varepsilon,s[i]})$. The variance $\sigma^2_{\varepsilon,s[i]}$ depends on the type of data system of observation i:

$$\varepsilon_i = 0 \qquad \text{if } s = \text{ CRVS}$$

$$\sigma_{\varepsilon,s} \sim \text{Normal}^{+}(0, 1^2) \text{ if } s = \text{health facility, HMIS, population study}$$

Note that the estimated variance for health facility > pop study > HMIS

Process model

$$\text{logit}(\phi_i) = \mu_i + \varepsilon_i$$

The ‘true’ transformed proportion μ_i is modeled as

$$\mu_i = \beta_0 + \beta_{r[i]} + \beta_{c[i]} + \beta_{p[i]} + \beta_{\text{NMR}} \log \text{NMR}_{c[i],t[i]} + \eta_{p[i],t[i]} + \gamma_{g[i],m[i]}$$

Hierarchical intercepts

$$\text{logit}(\phi_i) = \mu_i + \varepsilon_i$$

The ‘true’ transformed proportion μ_i is modeled as

$$\mu_i = \underbrace{\beta_0 + \beta_{r[i]} + \beta_{c[i]}} + \beta_{p[i]} + \beta_{\text{NMR}} \log \text{NMR}_{c[i],t[i]} + \eta_{p[i],t[i]} + \gamma_{g[i],m[i]}$$

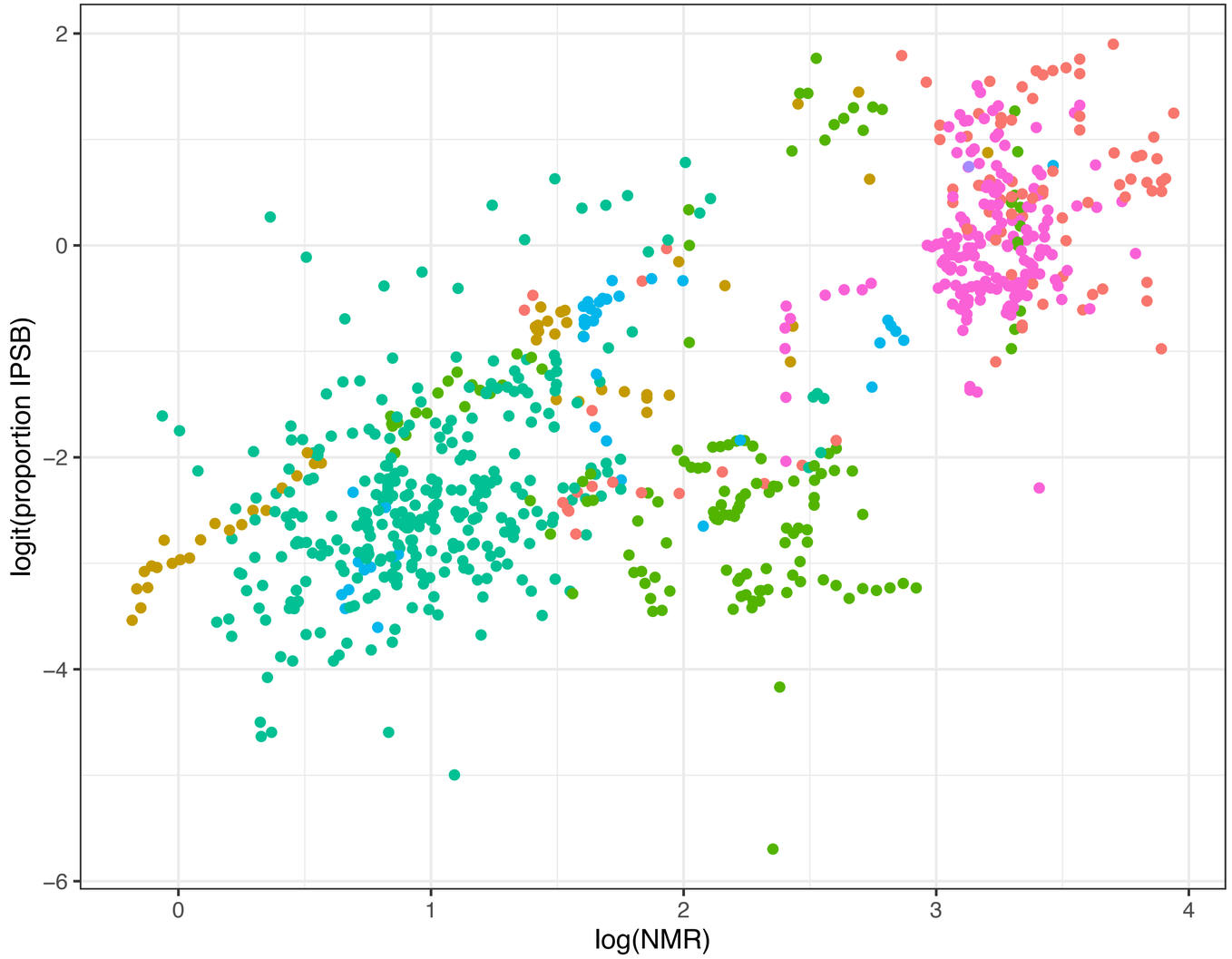
Model intercepts hierarchically (place within country within region within the world) to pool information across similar areas

Neonatal mortality as a covariate

$$\text{logit}(\phi_i) = \mu_i + \varepsilon_i$$

The ‘true’ transformed proportion μ_i is modeled as

$$\mu_i = \beta_0 + \beta_{r[i]} + \beta_{c[i]} + \beta_{p[i]} + \underbrace{\beta_{\text{NMR}} \log \text{NMR}_{c[i],t[i]}}_{\text{Trends in neonatal mortality rate inform trends in IPSB, allowing for reasonable trends in absence of data}} + \eta_{p[i],t[i]} + \gamma_{g[i],m[i]}$$


2
0
−2
−4
−6
logit(proportion IPSB)
0
1
2
3
4
log(NMR)
SDG Region
Central and Southern Asia
Eastern and South−Eastern Asia
Latin America and the Caribbean
North America, Europe, Australia and New Zealand
Northern Africa and Western Asia
Oceania (exc. Australia and New Zealand)
Sub−Saharan Africa

Penalized splines

$$\text{logit}(\phi_i) = \mu_i + \varepsilon_i$$

The ‘true’ transformed proportion μ_i is modeled as

$$\mu_i = \beta_0 + \beta_{r[i]} + \beta_{c[i]} + \beta_{p[i]} + \beta_{\text{NMR}} \log \text{NMR}_{c[i],t[i]} + \underbrace{\eta_{p[i],t[i]}} + \gamma_{g[i],m[i]}$$

Penalized splines component allows for data-driven trends in presence of reliable data. Spline coefficients modeled as random walk to ensure smoothness

Penalized splines

- To allow for data-driven trends we include a place-time specific component $\eta_{p,t}$, which is modelled using a first-order penalized splines set up

$$\eta_{p,t} = \sum_{h=1}^{H} k_h(t)\alpha_{h,p}$$

- Cubic B-splines $k_h(t)$ with knots placed at integer year values

- First-order differences in the coefficients $\Delta_{h,p}$are penalized to ensure a level of smoothness in the resulting fit:

$$\Delta_{h,p} = \alpha_{h,p} - \alpha_{h-1,p}$$

$$\Delta \sim \text{Normal}(0,\sigma_{\Delta}^2)$$

- Coefficients $\alpha_{.p}$ are constrained to sum to zero to ensure identifiability

Definitional adjustment

$$\text{logit}(\phi_i) = \mu_i + \varepsilon_i$$

The ‘true’ transformed proportion μ_i is modeled as

$$\mu_i = \beta_0 + \beta_{r[i]} + \beta_{c[i]} + \beta_{p[i]} + \beta_{\text{NMR}} \log \text{NMR}_{c[i],t[i]} + \eta_{p[i],t[i]} + \underbrace{\gamma_{g[i],m[i]}}$$

Definitional adjustment to account for different stillbirth definitions

Definitional adjustment

- Adjustment $\gamma_{g,m}$ for definition g (early or late) and income group m (high or low)
- Make use of auxiliary data which gives information on stillbirths by timing at different gestational ages:
    1. Euro-Peristat: high-quality data for 17 European countries (BUT country names are suppressed)
    2. Global Network Maternal Newborn Health Registry: information for 8 low- and middle-income countries
- Use overlapping data to inform prior distribution on adjustment term $\gamma_{g,m}$

Constructing country-level estimates

- Estimation of IPSB ϕ happens at ‘place’ level
- How to get country-level estimates $\hat{\phi}_{c,t}$?
- If the ‘place’ is just the country then

$$\hat{\phi}_{c,t} = \hat{\phi}_{p,t}$$

- But usually a ‘place’ is a subset of the whole country

Constructing country-level estimates

- If we had full coverage, and we knew the place weights w_p, then the country estimate would just be

$$\hat{\phi}_{c,t} = \sum_{p:c(p)=c} w_p \hat{\phi}_{p,t}$$

- …but in practice, we don't have full coverage, and we don't know place-specific weights

Constructing country-level estimates

Proposed weighting scheme:

$$\hat{\phi}_{c,t} = \sum_{p:c[p]=c} \hat{w}_p \hat{\phi}^{\text{obs}}_{p,t} + \left(1 - \sum_{p:c[p]=c} \hat{w}_p\right) \hat{\phi}^{\text{unobs}}_{c,t}$$

Constructing country-level estimates

Proposed weighting scheme:

$$\hat{\phi}_{c,t} = \sum_{p:c[p]=c} \hat{w}_p \hat{\phi}^{\text{obs}}_{p,t} + \left(1 - \sum_{p:c[p]=c} \hat{w}_p\right) \hat{\phi}^{\text{unobs}}_{c,t}$$

Weights are estimated based on ratio of observed number of stillbirths to estimated total stillbirths (with uncertainty)

Constructing country-level estimates

Proposed weighting scheme:

$$\hat{\phi}_{c,t} = \sum_{p:c[p]=c} \hat{w}_p \hat{\phi}^{\text{obs}}_{p,t} + \underbrace{\left(1 - \sum_{p:c[p]=c} \hat{w}_p\right) \hat{\phi}^{\text{unobs}}_{c,t}}$$

Weights are estimated based on ratio of observed number of stillbirths to estimated total stillbirths (with uncertainty)

Unobserved component accounts for under coverage

Constructing country-level estimates

Proposed weighting scheme:

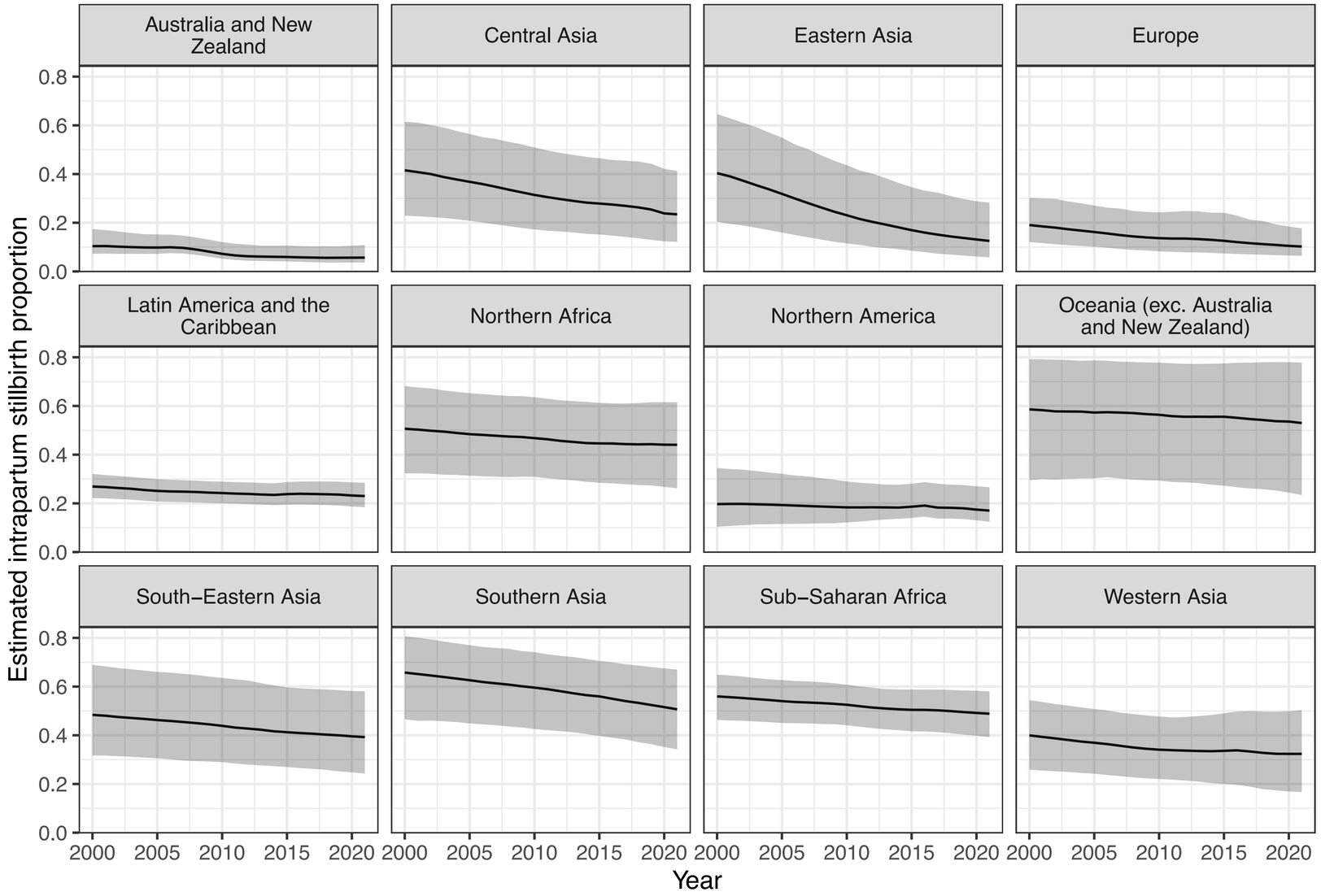
$$\hat{\phi}_{c,t} = \sum_{p:c[p]=c} \hat{w}_p \hat{\phi}^{\text{obs}}_{p,t} + \underbrace{\left(1 - \sum_{p:c[p]=c} \hat{w}_p\right)}_{\text{Unobserved component accounts for under coverage}} \hat{\phi}^{\text{unobs}}_{c,t}$$

Weights are estimated based on ratio of observed number of stillbirths to estimated total stillbirths

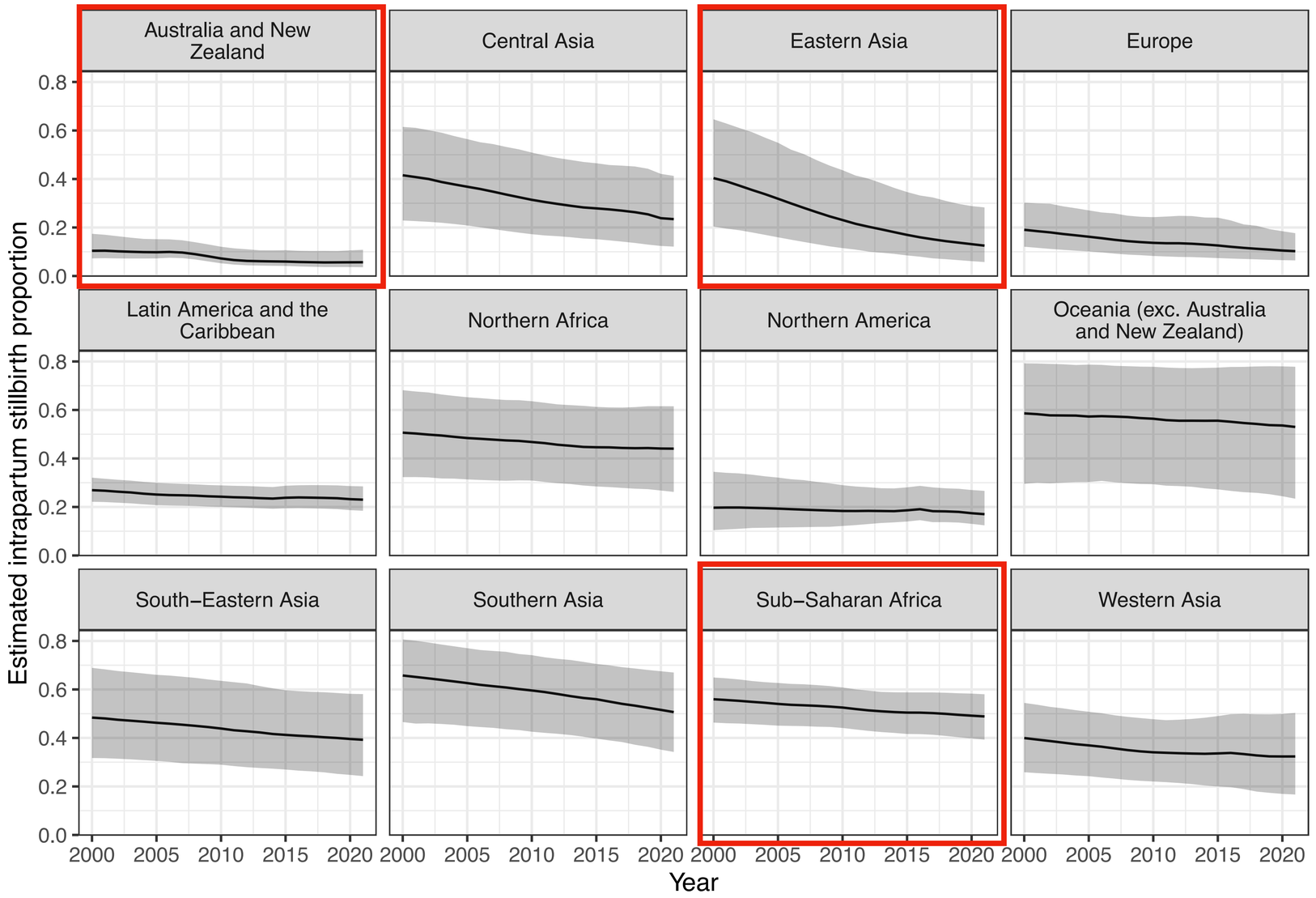
Informed by NMR and regional patterns

Illustrative Results

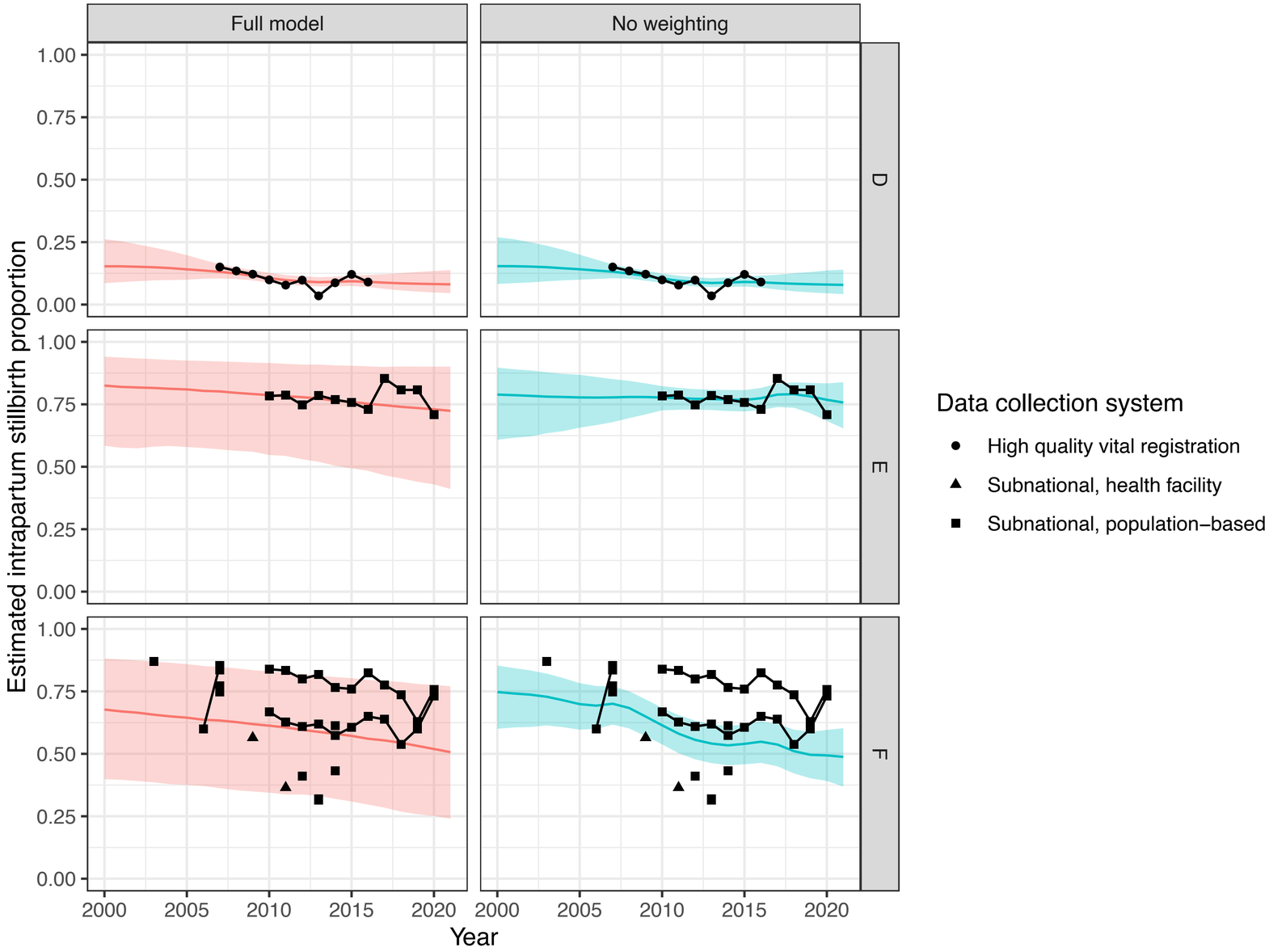
Region estimates



Region estimates



Impact of weighting



Summary

- Proportion of stillbirths that are intrapartum is an important indicator to monitor to track progress towards the goal of ending preventable stillbirths
- Data availability and quality varies substantially by region, with other definition and classification differences also reducing comparability
- Bayesian hierarchical penalized splines model with post weighting accounts for many of issues, and performs reasonably well in a series of validations
- Future work: improve estimation of weights; estimation of total stillbirths and timing in one model

Thanks!

monica.alexander@utoronto.ca

monicaalexander.com

 @monjalexander

 MJAlexander

Gestational adjustment

- Let $\dot{y}_{c,g}$and $\dot{z}_{c,g}$ denote respectively intrapartum and antepartum stillbirth counts for some country c and some gestational definition g. The counts are modeled

$$\dot{y}_{c,g} \,|\, \rho_{c,g} \sim \text{Binomial}(\dot{y}_{c,g} + \dot{z}_{c,g}, \rho_{c,g})$$

$$\text{logit}\rho_{c,g} = \nu_{c,g} + \gamma_{g,m[c]}$$

$$\nu_{c,g} \sim \text{Normal}(0, 10^2)$$

- where $\nu_{c,g}$ is given a vague prior and represents the mean under the late gestational age definition. The difference between the proportions in the early and late definitions is therefore captured by the adjustment factor.

Construction of weights

- We construct an estimate $\hat{w}_p$ as the ratio of the number of observed classified stillbirths in place p to the number of total stillbirths expected nationally, based on UN IGME estimates of overall stillbirths.
- Let $s_i = y_i + z_i$ denote the sum of observed stillbirths classified as intrapartum or antepartum i.
- Let $\tilde{S}_i = \tilde{S}_{c[i],t[i]}$ denote the estimate of total stillbirths from the UN IGME total stillbirth rate model in the country c and year t corresponding to the observation.
- To reflect uncertainty in the number of stillbirths, we directly use posterior samples of $\tilde{S}_i$ when computing our own posterior samples.
- Estimated weights are

$$\hat{w}_p = \frac{\sum_{i:p[i]=p} s_i}{\sum_{i:p[i]=p} \tilde{S}_i}$$

Unobserved component

- The "unobserved" component for a country is centered at the estimate given its region and country intercepts and NMR level

$$\hat{\mu}_{c,t}^{\text{unobs}} = \hat{\beta}_0 + \hat{\beta}_{r[c]} + \hat{\beta}_c + \tilde{\beta}_{p_c} + \hat{\beta}_{\text{NMR}} \log \tilde{\text{NMR}}_{c,t} + \tilde{\eta}_{c,t}$$

$$\hat{\phi}_{c,t}^{\text{unobs}} = \text{logit}^{-1}(\hat{\mu}_{c,t}^{\text{unobs}})$$

- $\tilde{\beta}_{p_c}$ and $\tilde{\eta}_{c,t}$ are new realizations of the sub-population effect and time trend to reflect appropriate uncertainty about the unobserved population

Validation results

Table 4: Model evaluation metrics using 2000-2016 data as a training set and data from 2017 onward as a test set.

Region	Mean absolute error	95% prediction interval coverage
Global	0.044	0.917
Central and Southern Asia	0.095	0.850
Eastern and South-Eastern Asia	0.014	1.000
Latin America and the Caribbean	0.028	0.960
North America, Europe, Australia and New Zealand	0.028	0.922
Northern Africa and Western Asia	0.049	0.615
Oceania (exc. Australia and New Zealand)	0.271	1.000
Sub-Saharan Africa	0.052	0.979

Validation results

Table 5: Model evaluation metrics from 10-fold cross validation.

Region	Mean absolute error	95% prediction interval coverage
Global	0.041	0.925
Central and Southern Asia	0.102	0.943
Eastern and South-Eastern Asia	0.039	0.980
Latin America and the Caribbean	0.025	0.928
North America, Europe, Australia and New Zealand	0.021	0.927
Northern Africa and Western Asia	0.029	0.829
Oceania (exc. Australia and New Zealand)	0.271	1.000
Sub-Saharan Africa	0.063	0.916

UNIVERSITY OF MICHIGAN

Michigan Statistics for Individualized Healthcare Lab

MiSIL

Developing Innovative Statistical Framework to Integrate Multiple Verbal Autopsy Datasets to Estimate Cause-Specific Mortality Fractions

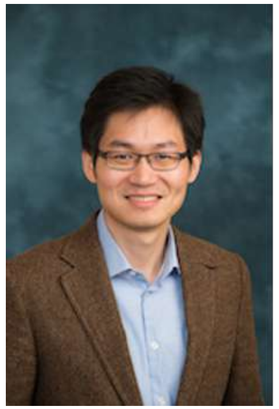
IAOS-ISI 2024 Mexico City

Zhenke Wu, PhD

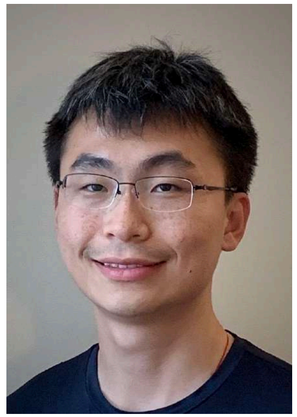
Associate Professor of Biostatistics, University of Michigan
Michigan Institute for Data Science (MIDAS)
Michigan Statistics for Individualized healthcare Lab (MiSIL)

some slides sourced from Prof. Richard Li

Team



Zhenke Wu
Associate Professor
of Biostatistics, UMich



Zehang Richard Li
Assistant Professor
of Statistics,
UC Santa Cruz

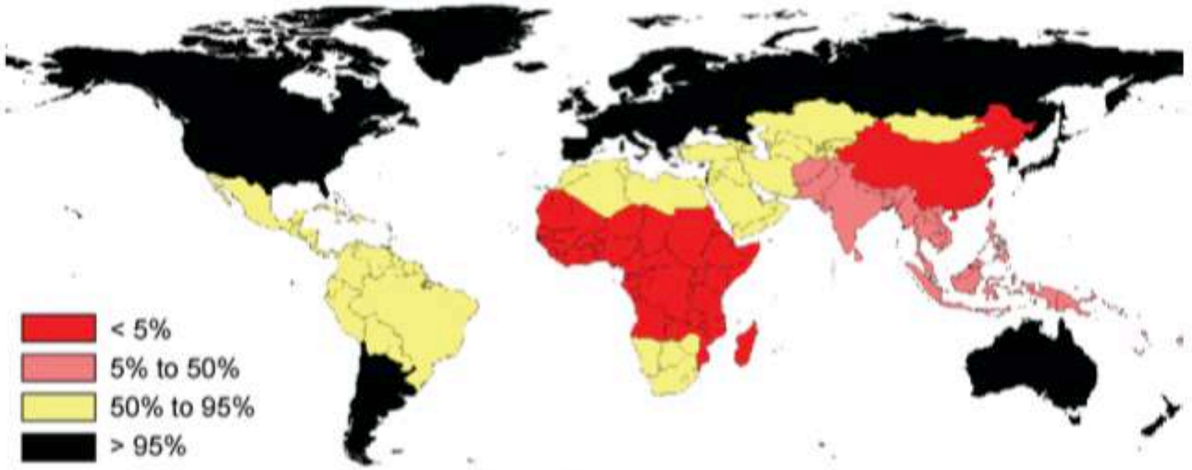
Former PhD Student
Biostatistics, UMich;
Currently at
Max Planck Institute for
Demographic Research, Germany

PhD Candidate
Biostatistics, UMich

Outline

- Part 1: Background, gaps, and challenges
- Part 2: Proposed Bayesian approach
    - Likelihood: nested latent class models
    - Prior (for integrating similarity info between multiple datasets, or "domains"; encoded by a tree)
- 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

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

Population Health Metrics Research Consortium Adult and Adolescent Verbal Autopsy Module 1

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

Population Health Metrics Research Consortium Adult and Adolescent Verbal Autopsy Module 13

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”

- analytic methods + reproducible open-source software —> 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, Li et al. (2024)
  - openva.net (Clark, McCormick, Li and others): dedicated to open-sourcing stat tools for VA research

Li et al. (2023). The openVA toolkit for verbal autopsies. *The R Journal*

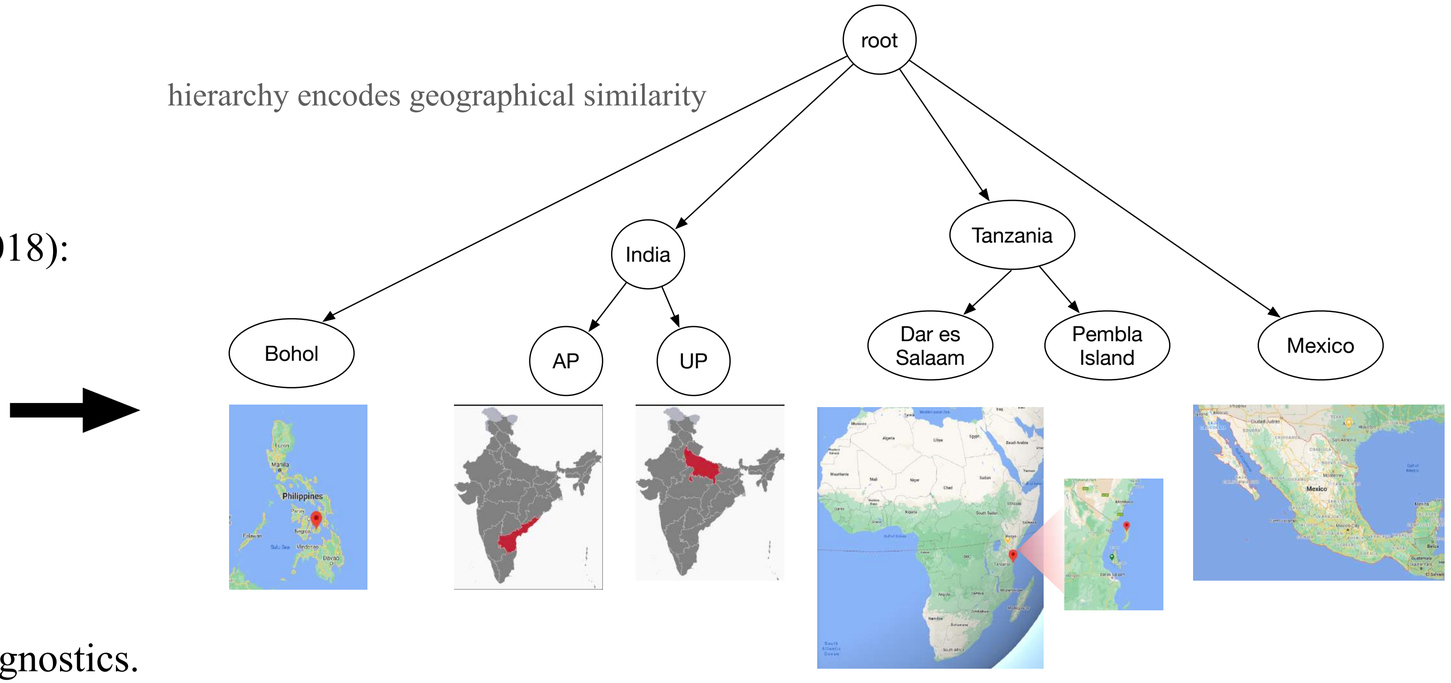
“The Pain of Growth”

- Expansion of VA to new “domains”: new regions (e.g., Brazil, New Guinea) and/or new time periods (COVID vs non-COVID periods)
  - 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

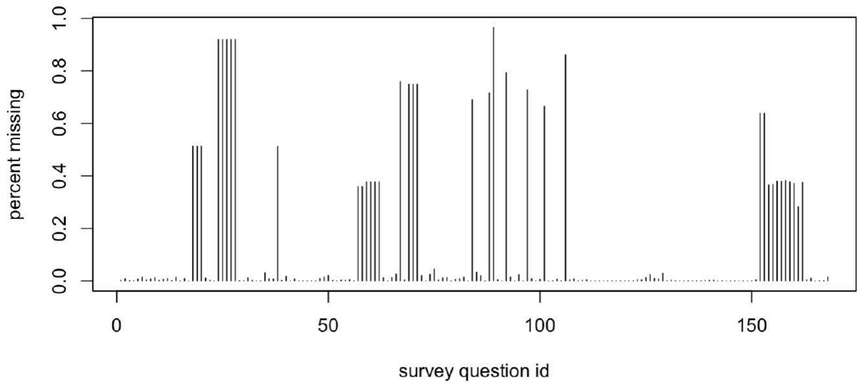
• The PHMRC VA gold-standard data (Population Health Metrics Research Consortium, 2018):

- 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 focus on 35 CODs at the finest level and 168 binary indicators.
  - N = 7841;
  - Differential rates of missingess (see right figure: "Don't Know", "Refused to Answer" and no data.)

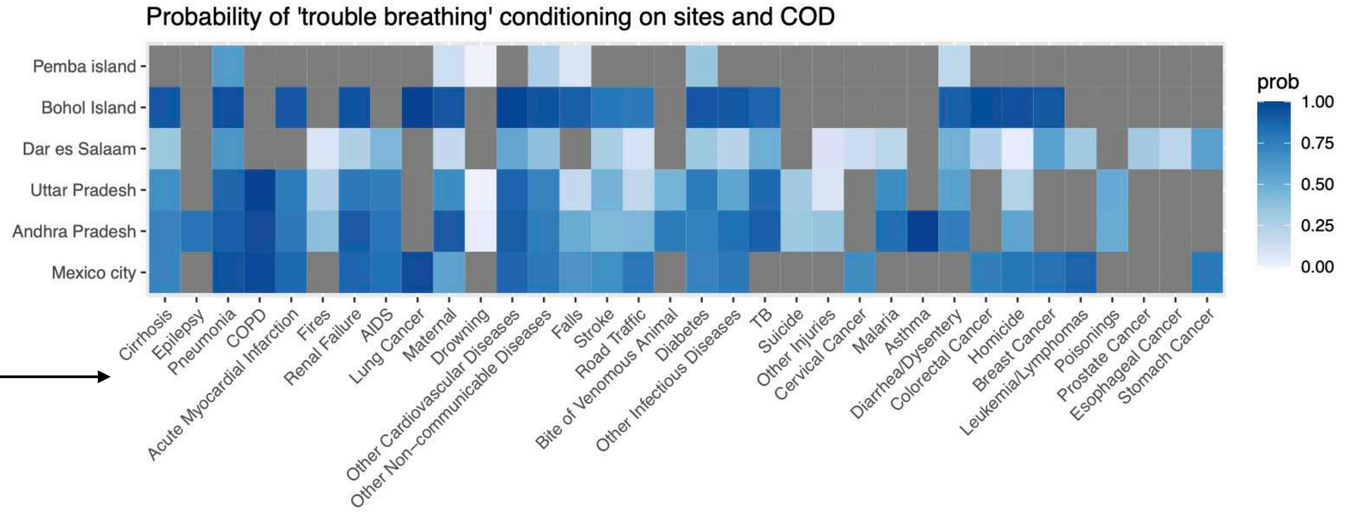


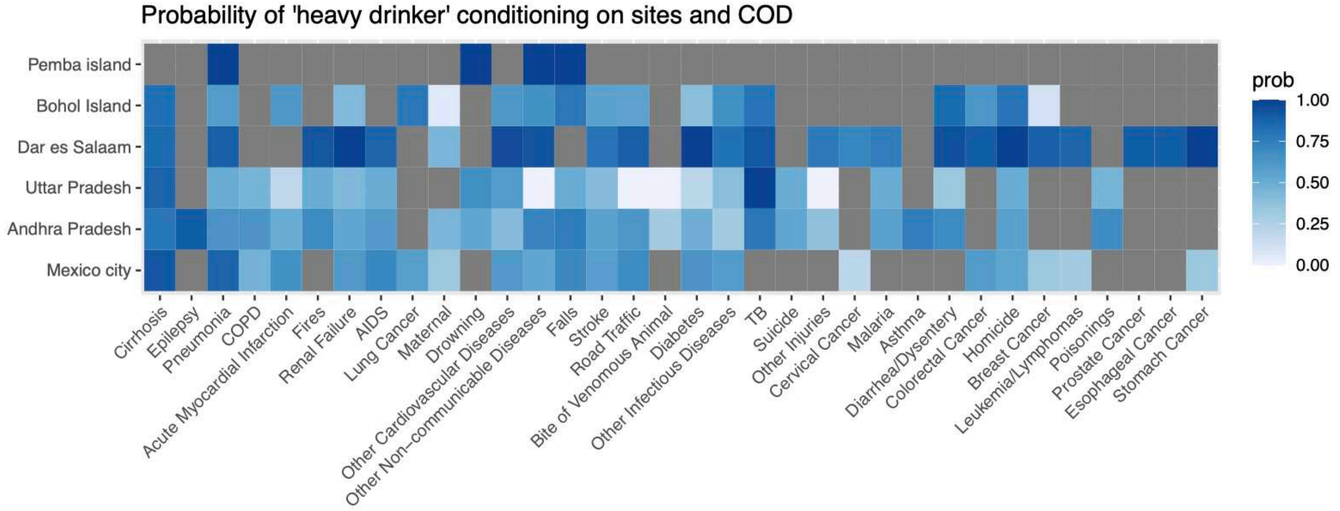
• We will take one site as the target and use the other five sites as source domains.

Example of Between-Domain Differences: PHMRC Data

domains →

causes →





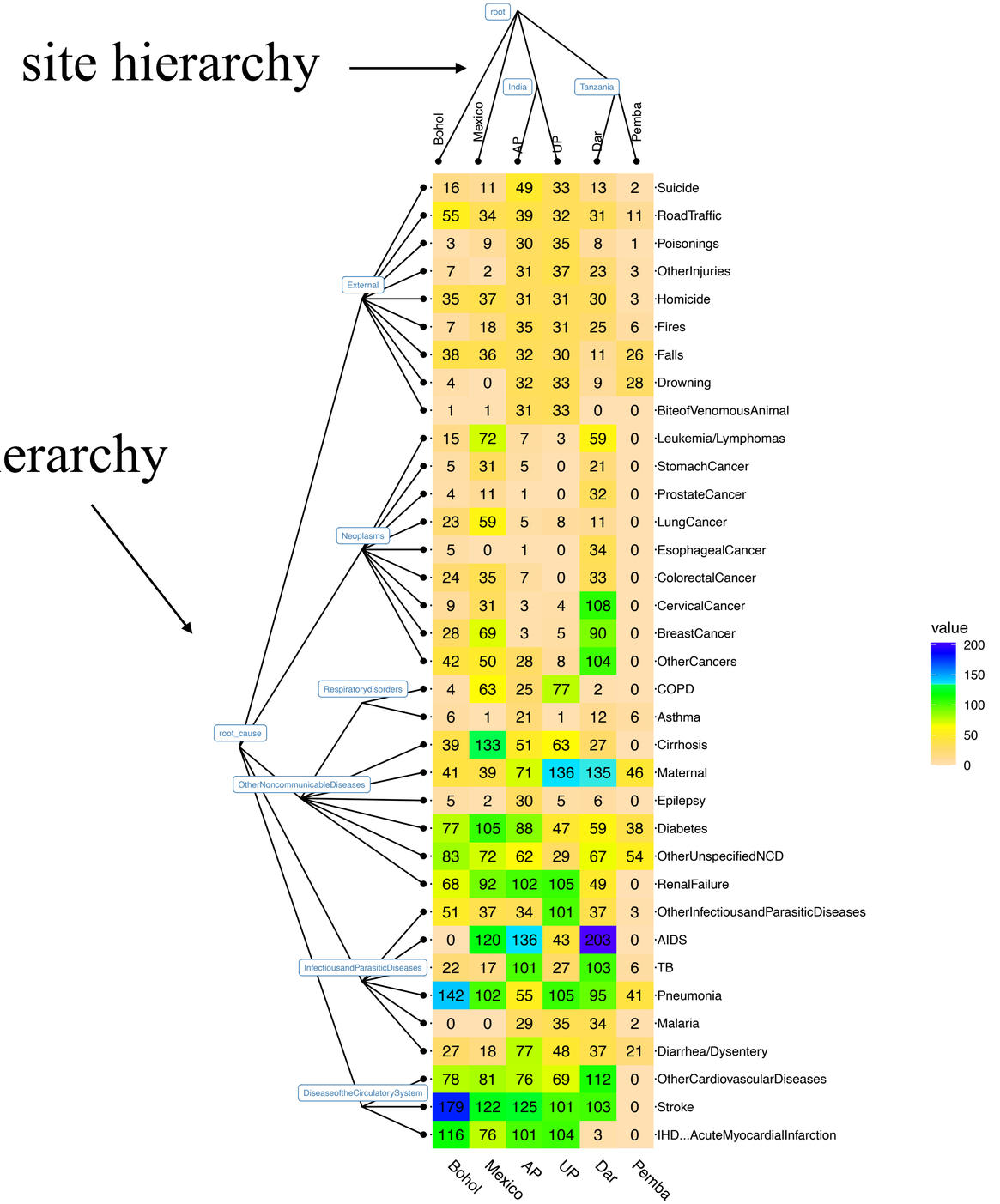
Plots including only symptom–cause pairs with least 20 observations.

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.

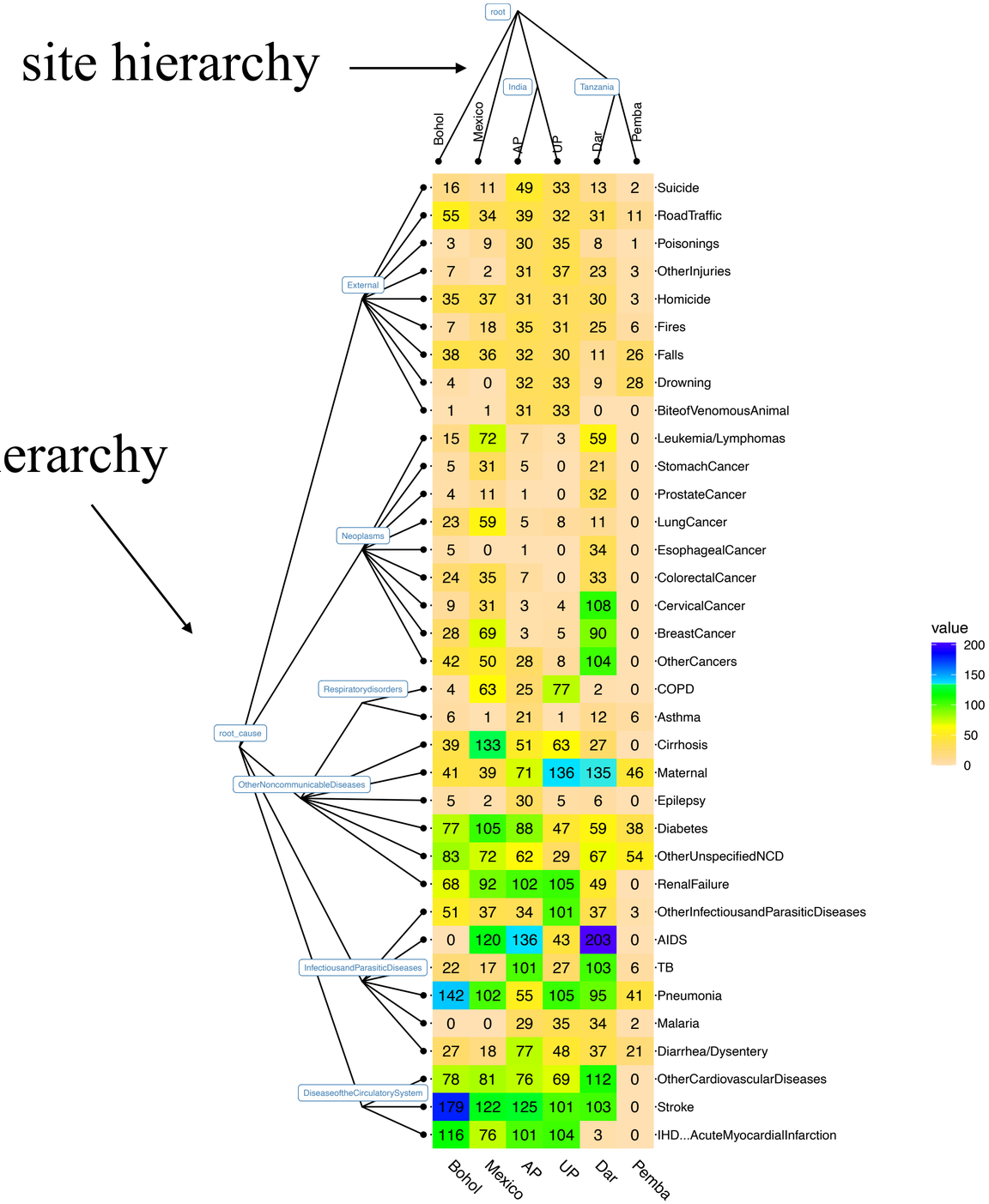
The exact death counts are shown in corresponding cells.



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



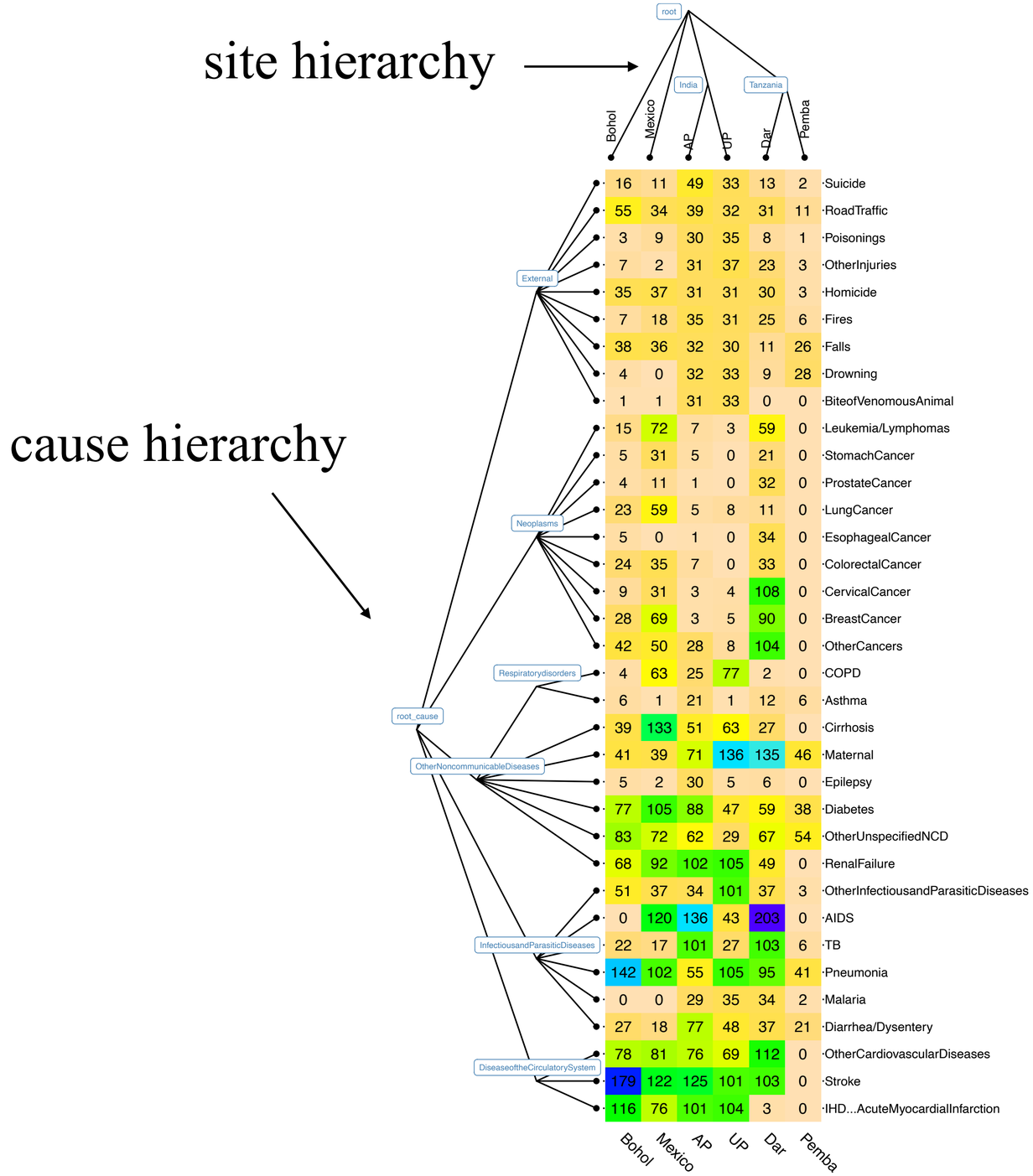
Data: A Closer Look

Death Counts

In our experiments:

Mask the causes-of-death in one site (column)

- **target** domain: masked site
- source domains: the rest sites



Notation

- $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iJ})^\intercal \in \{0,1\}^J$: 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.

Our Framework: Nested Latent Class Models

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

$$\text{cause of death}: \quad Y_i \mid D_i = g \sim \text{Categorical}_C(\boldsymbol{\pi}^{(g)}), \tag{1}$$

$$\text{latent class}: \quad Z_i \mid Y_i = c, D_i = g \sim \text{Categorical}_K(\boldsymbol{\lambda}^{(c,g)}), \tag{2}$$

$$\text{responses}: \quad X_{ij} \mid Z_i = k, Y_i = c \overset{\text{indep.}}{\sim} \text{Bernoulli}(\theta_{jk}^{(c)}), j \in [J] \tag{3}$$

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 $\{Y_i : D_i = 0\}$.

Why bother?

- Given each cause, the conditional distribution of symptom approximated by a latent class model. 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

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”

Prior Distribution to Integrate the Tree Information

Condensed Summary

- Tree-informed Bayesian shrinkage prior
  - heuristics: “parameters connected by shorter paths in a tree *a priori* take more similar values”
  - we apply this framework to let $\lambda^{(c,u)}$ parameters diffuse along the tree (domain hierarchy)

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

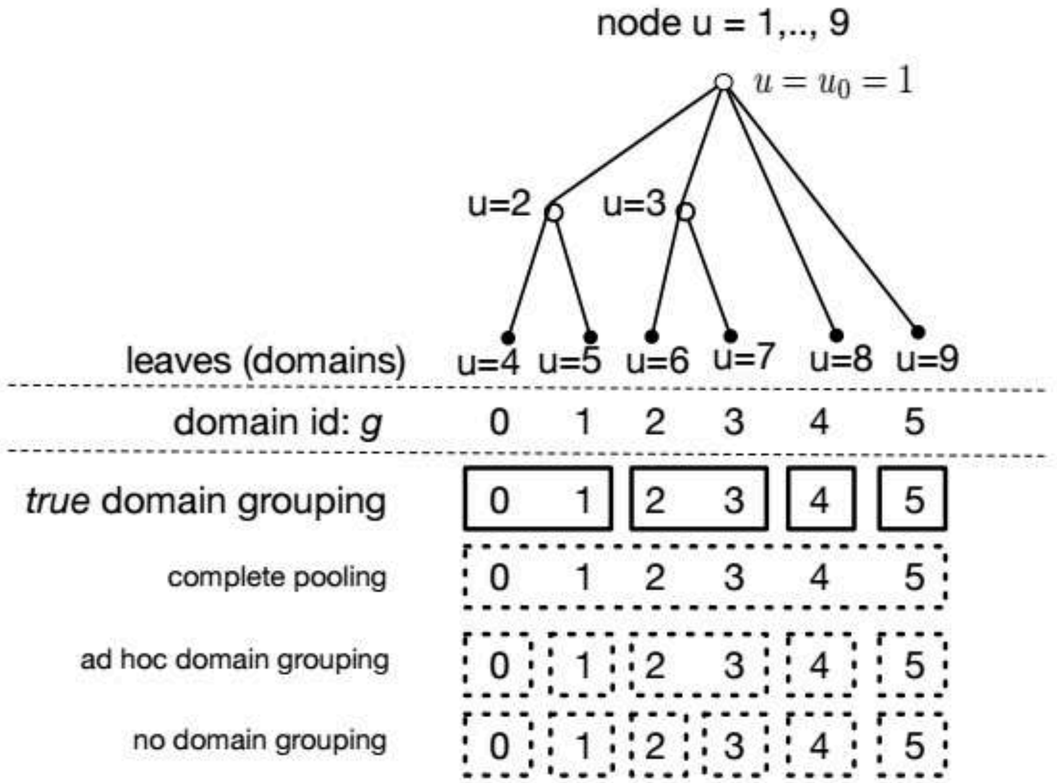
Results

- Performance Metrics
  - CSMF accuracy: normalized L1 distance

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

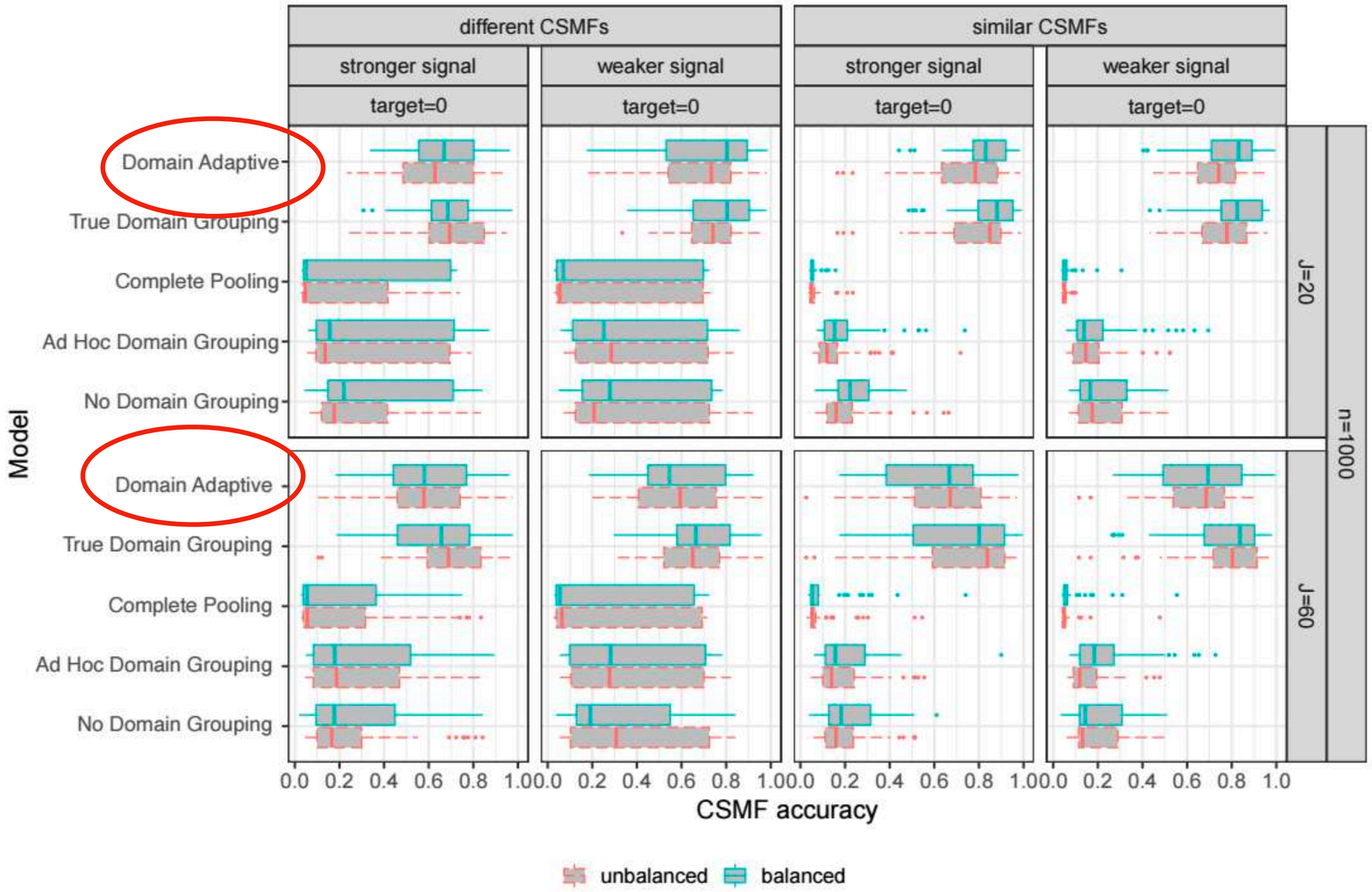
- Top cause accuracy, top 3 cause accuracy (McCormick et al. 2016 JASA)
  - CSMF accuracy (Murray et al. 2011, Population Health Metrics)

Simulation Results



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

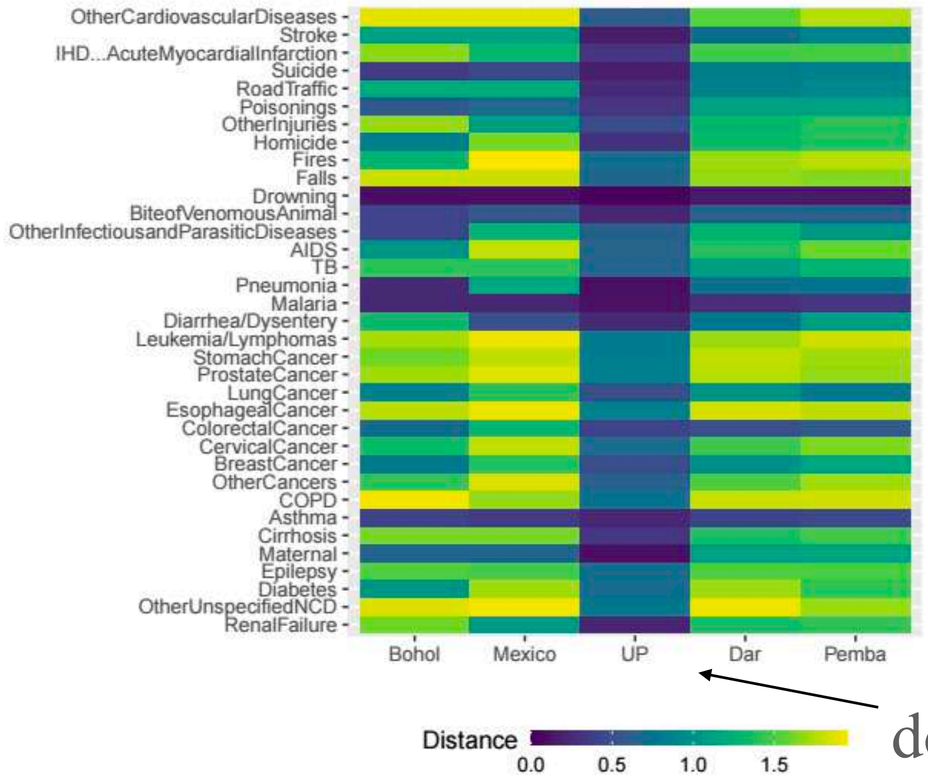
Simulation Results



(b) Simulation I: CSMF accuracy comparison.

PHMRC Data Results: “Similarity”

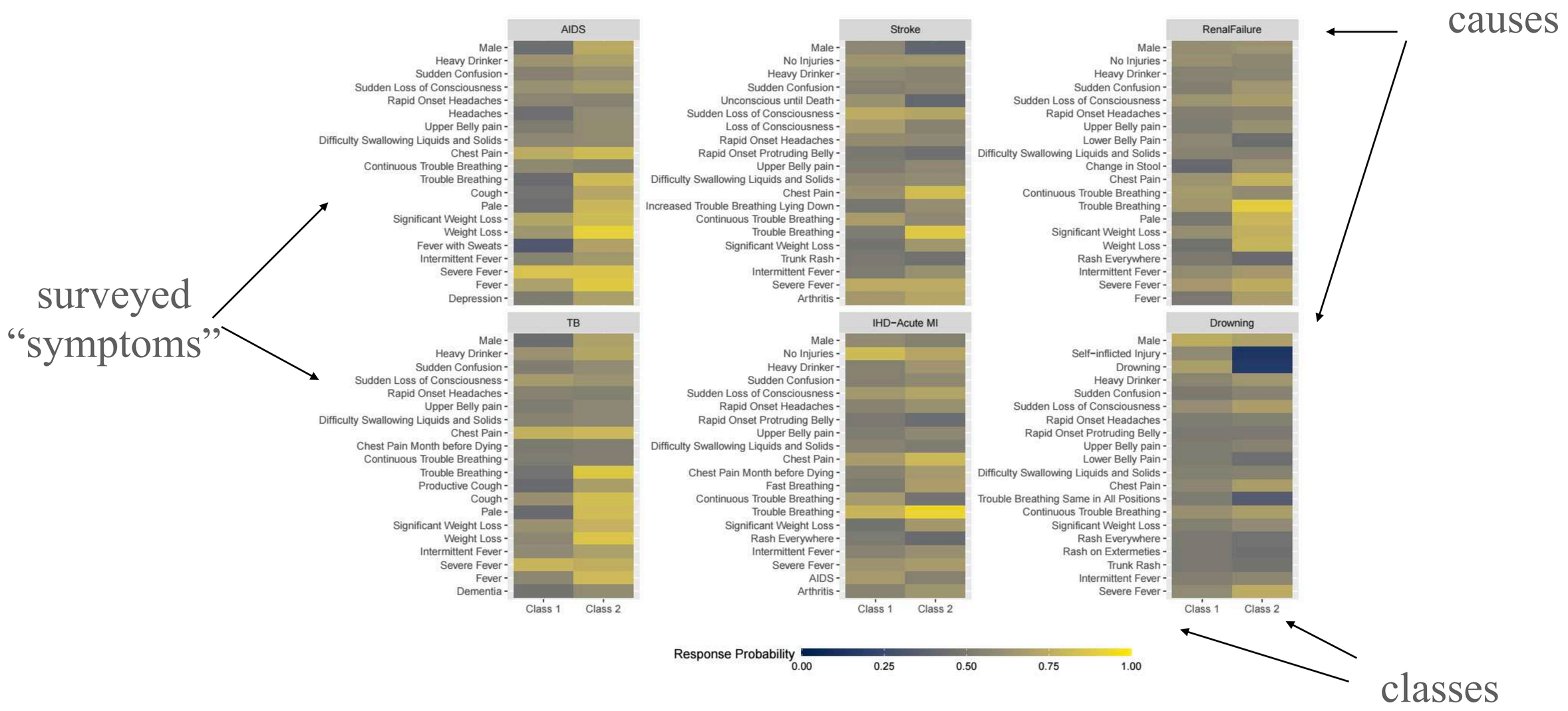
causes



domains
(all but “AP”, a site from India)

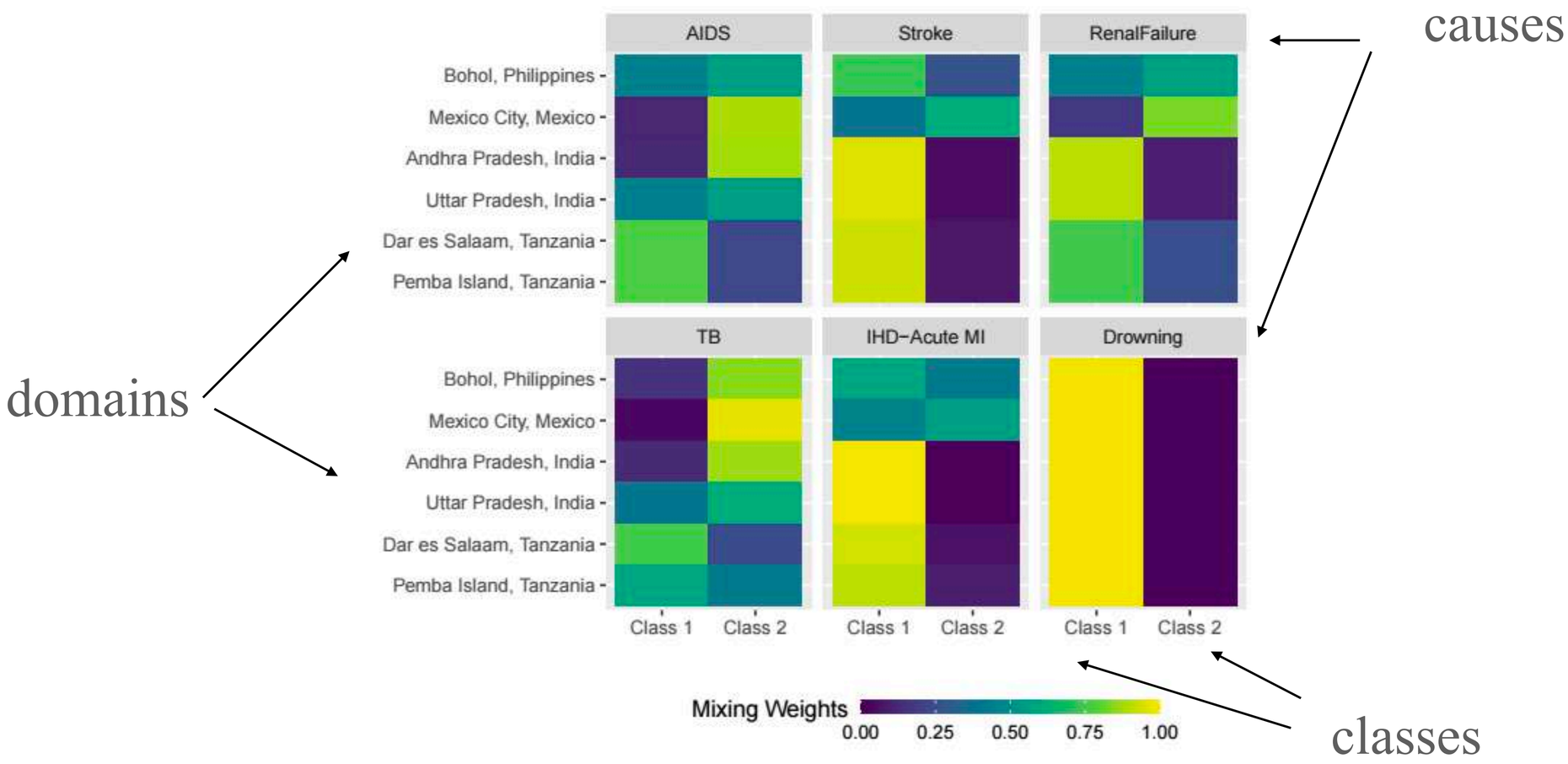
(c) Estimated cause-specific cophenetic distances between AP (target) and each of the five source domains; 35 rows representing 35 causes used during model fitting.

PHMRC Data Results: “class profiles”



(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: "For some causes, domains differ in how the classes got mixed"



(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

Methods

- Current work assumed the 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.
- **A further study of how to deal with cause-of-death labeled at multiple resolutions**

Future Directions

Applied

- How to deal with emerging prominent causes over different time periods (COVID19…)?
- How to actively choose the most informative deaths to label?
- 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. (2024). Tree-informed Bayesian multi-source domain adaptation: cross-population probabilistic cause-of-death assignment using verbal autopsy. Biostatistics. https://doi.org/10.1093/biostatistics/kxae005

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