Methods To Measure All-Cause and Cause-Specific Mortality in Places Without Vital Statistics Systems

Samuel J. Clark

Department of Sociology, The Ohio State University, Columbus, USA

UT Austin PRC/Sociology Seminar

February 10, 2025
Austin, Texas

Overview

Introduction

Example 1: Age-specific Mortality Models

Example 2: Verbal Autopsy

Closing

About me

Demographer → Epidemiologist, Data Scientist/Statistician

How did I get here?

- Born in **Kenya** and grew up in **East Africa**
- Moved to USA at age 16
- Undergraduate at Caltech
  - BS Biology (neurobiology)
  - BS Engineering (computer science, electrical engineering)
- Graduate work at University of Pennsylvania
  - PhD in demography
  - Field work in Zambia
- Postdoc in South Africa
  - 5 years working with health and demographic surveillance systems, lived in Durban
- About 10 years working with statisticians and UN Population Division on methods

Mortality

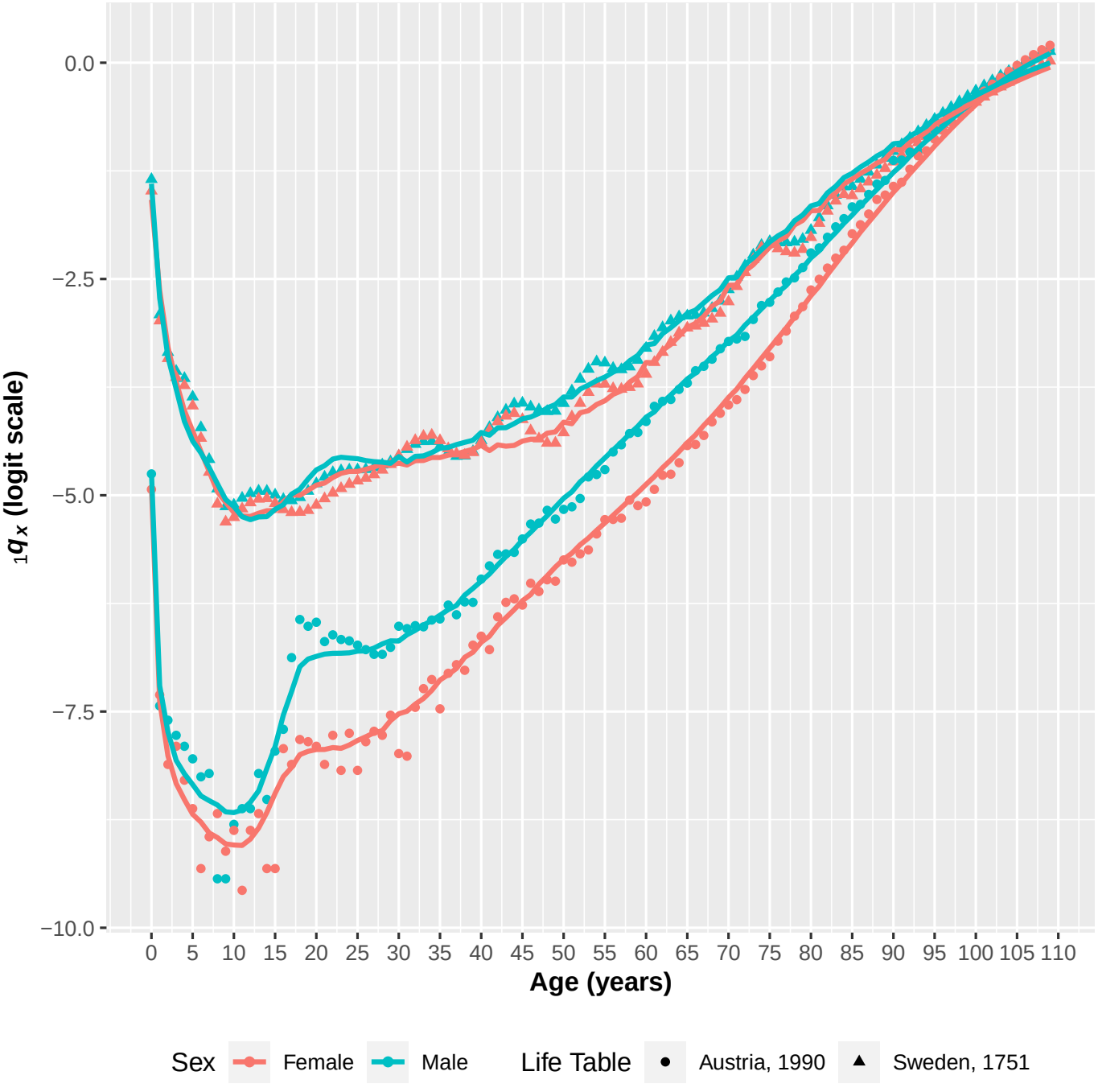
Age-specific mortality

- Human mortality varies systematically by age – very young and old have higher mortality
- Overall, mortality has decreased dramatically in the past century or so
- Mortality measured as either
  - Rate: ${}_nM_x = \frac{\text{deaths}}{\text{person-years}}$, often log-transformed to not be bounded by 0
  - Probability: ${}_nq_x = \frac{\text{deaths age x} \rightarrow \text{x+n}}{\text{alive at age x}}$, often logit-transformed to not be bounded by 0 or 1: $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$

Why do we care?

- Sensitive reflection of population health
- Actuarial sciences – predict mortality
- Epidemiology and demographic models: e.g. population health and population forecasting

Example logit-transformed age schedules of mortality: ${}_1q_x$



Age-specific mortality model using the SVD

(Good, 1969; Stewart, 1993; Strang, 2009)

The singular value decomposition (SVD) of a generic matrix $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{USV}^{\mathsf{T}} \tag{1}$$

- By construction, the first RSV points in the direction that captures the greatest possible variation in the cloud of points, and subsequent RSVs sequentially capture as much of the remaining variation as possible

Equation 1 can be rearranged so that each column, $\mathbf{x}_\ell$, of $\mathbf{X}$ is represented as the weighted sum of LSVs:

$$\mathbf{x}_\ell = \sum_{i=1}^{\rho} s_i v_{\ell i} \mathbf{u}_i \tag{2}$$

- The fact that the first RSV is associated with the direction of greatest variation in the cloud of points defined by $\mathbf{X}$ means that the first term in this sum accounts for the bulk of the variation among the columns $\mathbf{x}_\ell$ of $\mathbf{X}$, likewise the second RSV for what is left over, etc. (Golub et al., 1987)
- **In general, a small number of terms is sufficient to closely approximate $\mathbf{x}_\ell$**

SVD mortality model

Data: large, varied dataset of sex-, age-specific mortality

- Human Mortality Database (HMD): $\sim$ 10,000 age-specific mortality schedules
- All high-quality mortality data spanning past 200 years from countries with accurate death reporting, i.e. rich countries!

Calibration

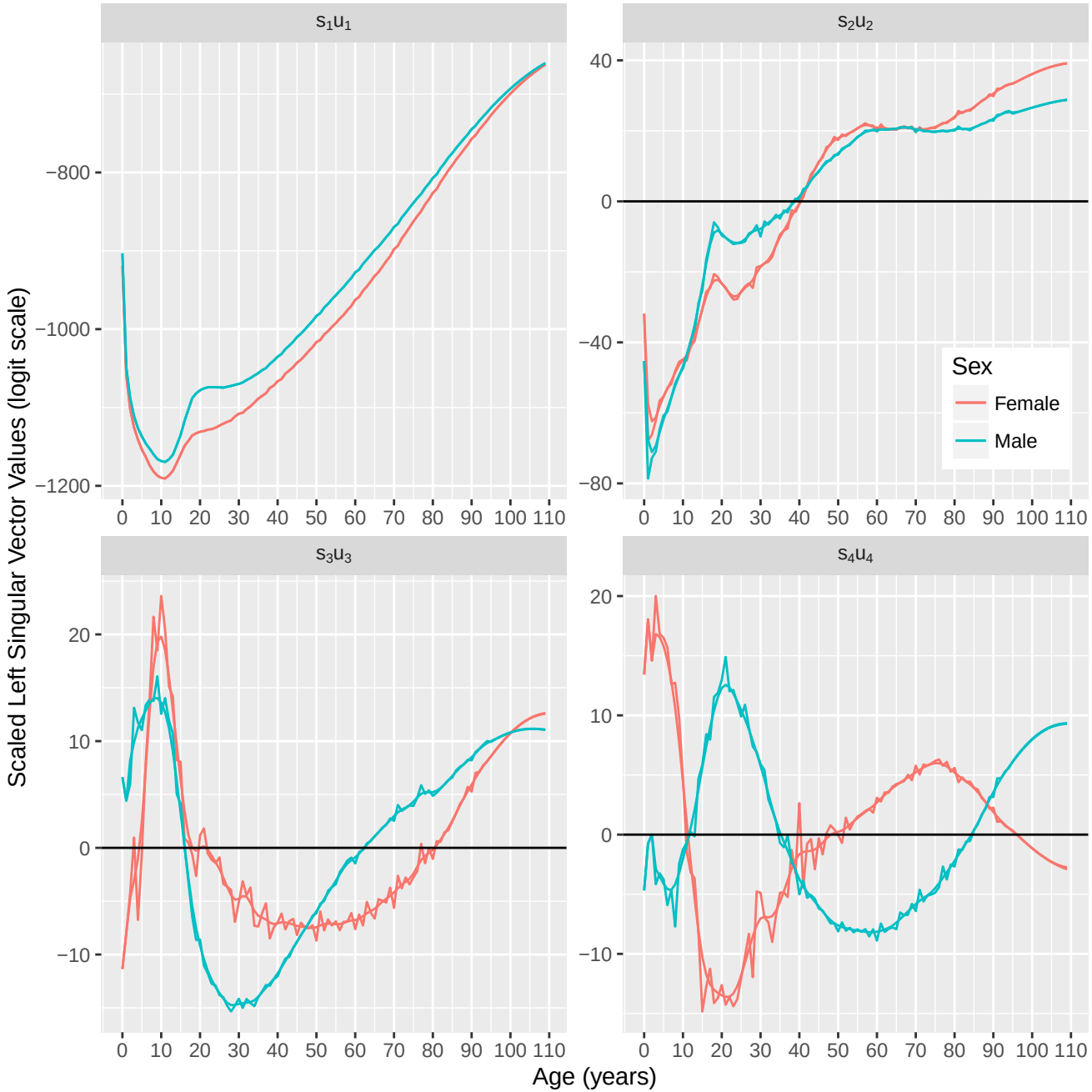
- SVD of HMD yields
  - Constant age-varying components - LSVs that represent systematic shape of human mortality by age
  - Varying weights - RSVs whose elements are the weights necessary in Equation 2 to fully reconstruct the original HMD mortality schedules

Prediction

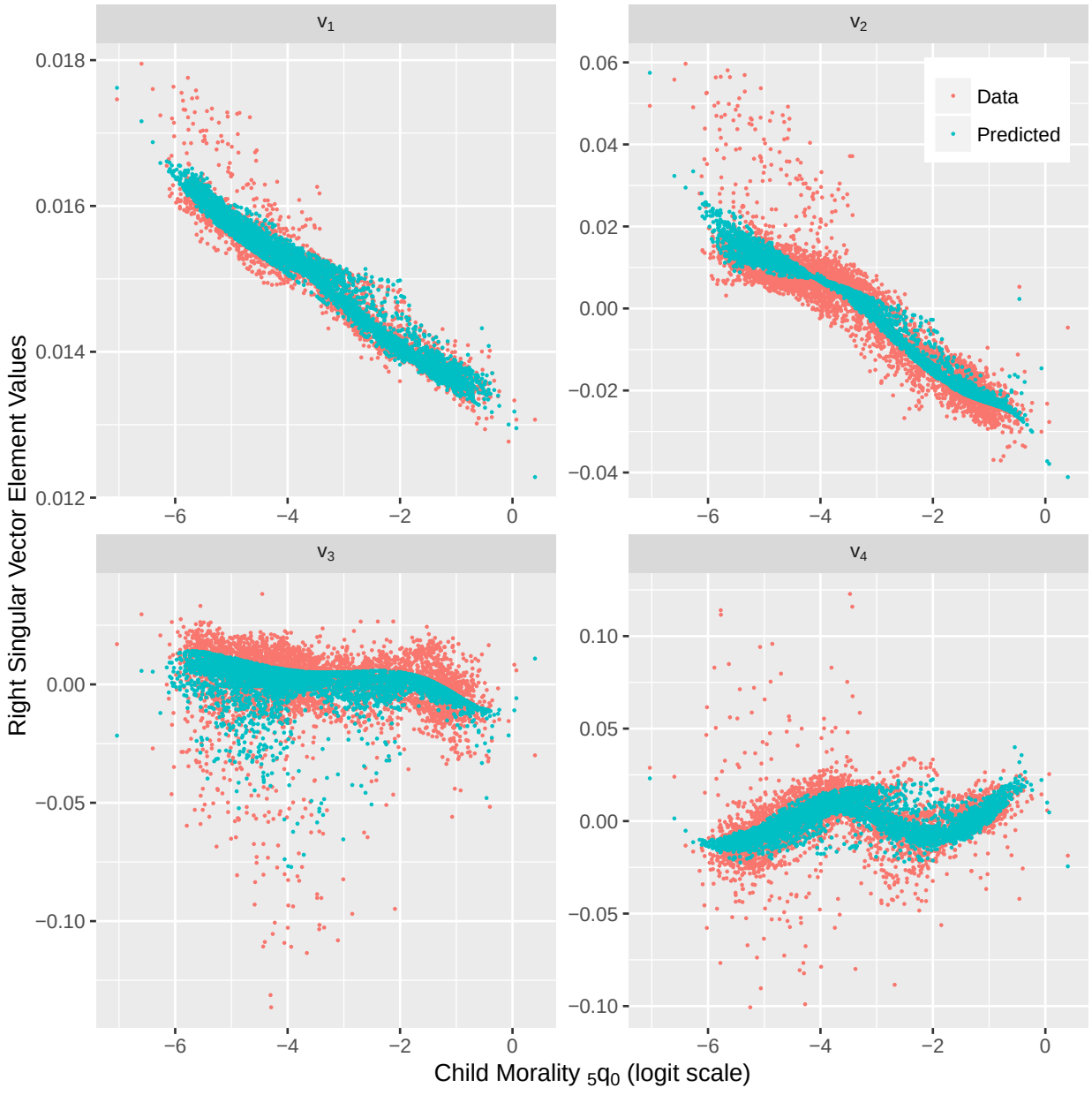
- Build statistical models that relate weights (RSV elements) to interesting predictor that varies with human mortality
- Use those models to predict weights as functions of the predictor, and use the weights to reconstruct full schedules of age-specific mortality

Super parsimonious, efficient model of 110 1-year age-specific mortality values

Scaled left singular vectors **u** (LSV) from SVD of HMD



Relationship between right singular vector **v** (RSV) elements (weights) and level of child mortality



Application: UN Population Division HIV model

World Population Prospects (Link)

- Produced by UN Population Division
- Biannual population estimates and forecasts for all countries of the world
- Requires models for age-specific mortality
  - Estimates for countries with poor data
  - Forecasts for all countries, especially those with poor data
- Special need for countries with endemic HIV – specific age pattern associated with prevalence of untreated HIV

My team's input

- Created SVD-based component model of age-specific mortality calibrated with mortality data from HMD and HIV-affected countries
- Predicts 1-year age group mortality as function of HIV prevalence and ART coverage
- Created package for the `R` statistical programming environment that implements the model
- Model used by UN Population Division for WPP 2022 and 2024
- Developing refinements and model for all countries for WPP 2026

Component model materials

- Paper: A General Age-Specific Mortality Model With an Example Indexed by Child Mortality or Both Child and Adult Mortality. *Demography*, 2019.
- Reproducibility materials: `http://github.com/sinafala/svd-comp`
- SVDMx `R` package on CRAN: https://cran.r-project.org/package=SVDMx

DEMOGRAPHY

ISSUES FEATURED ADVANCE PUBLICATION FOR AUTHORS ALERTS ABOUT

Volume 56, Issue 3
June 1, 2019

Previous Issue Next Issue

Previous Article Next Article

RESEARCH ARTICLE | MAY 28 2019

A General Age-Specific Mortality Model With an Example Indexed by Child Mortality or Both Child and Adult Mortality

Samuel J. Clark

Demography (2019) 56 (3): 1131–1159.
https://doi.org/10.1007/s13524-019-00785-3

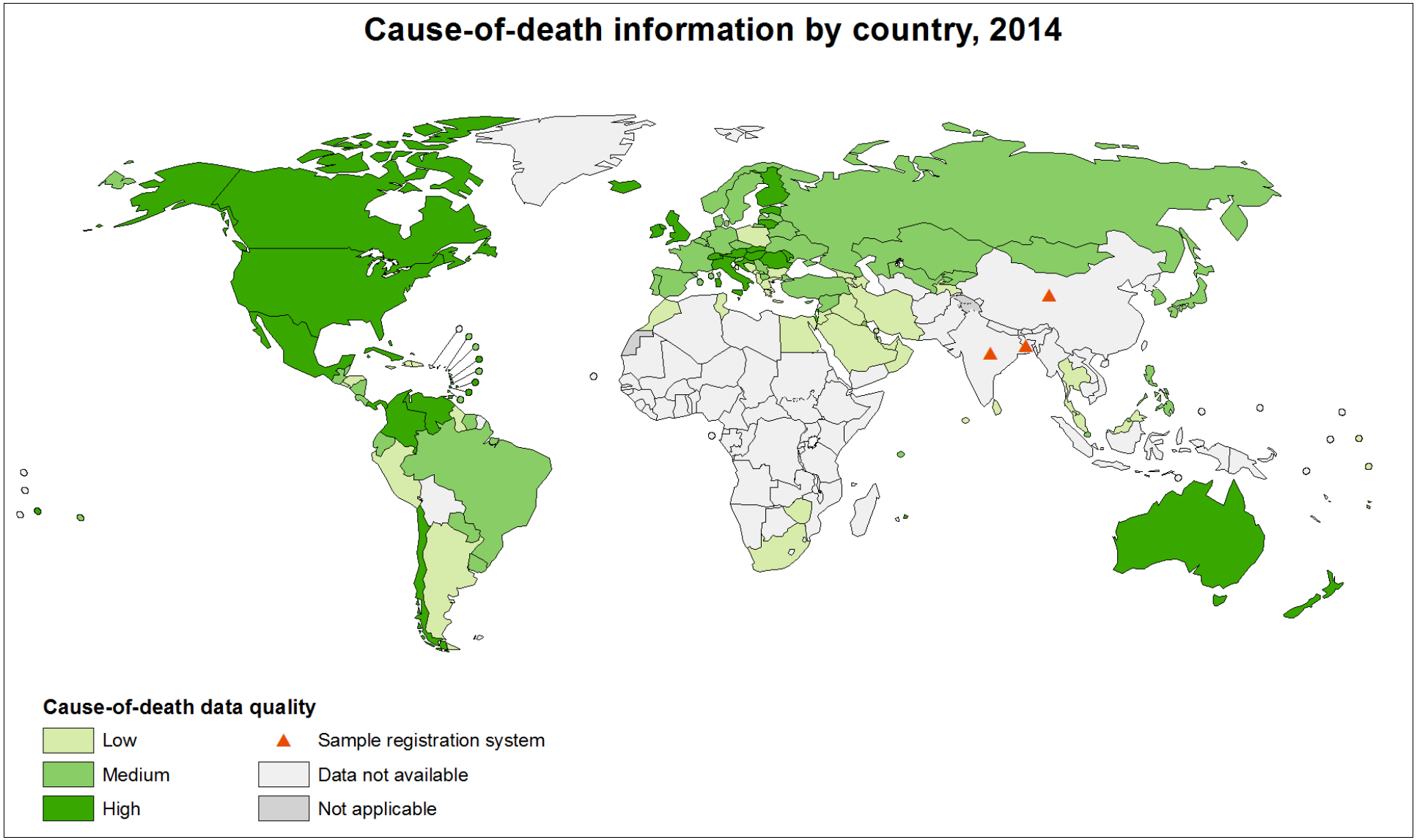
Share Tools

Abstract

The majority of countries in Africa and nearly one-third of all countries require mortality models to infer the complete age schedules of mortality that are required to conduct population estimates, projections/forecasts, and other tasks in demography and epidemiology. Models that relate child mortality to mortality at other ages are important because almost all countries have measures of child mortality. A general, parameterizable component model (SVD-Comp) of mortality is defined using the singular value decomposition and calibrated to the relationship between child or child/adult mortality and mortality at other ages in the observed mortality schedules of the Human Mortality Database. Cross-validation is used to validate the model, and the predictive performance of the model is compared with that of the log-quadratic (Log-Quad) model, which is designed to do the same thing. Prediction and cross-validation tests indicate that the child mortality–calibrated SVD-Comp is able to accurately represent the observed mortality schedules in the Human Mortality Database, is robust to the selection of mortality schedules used for calibration, and performs better than the Log-Quad model. The child mortality–calibrated SVD-Comp can be used where and when child mortality is available but mortality at other ages is unknown.

Global cause of death information – **burden of disease**

(Nichols et al., 2018)



Verbal autopsy – **VA**

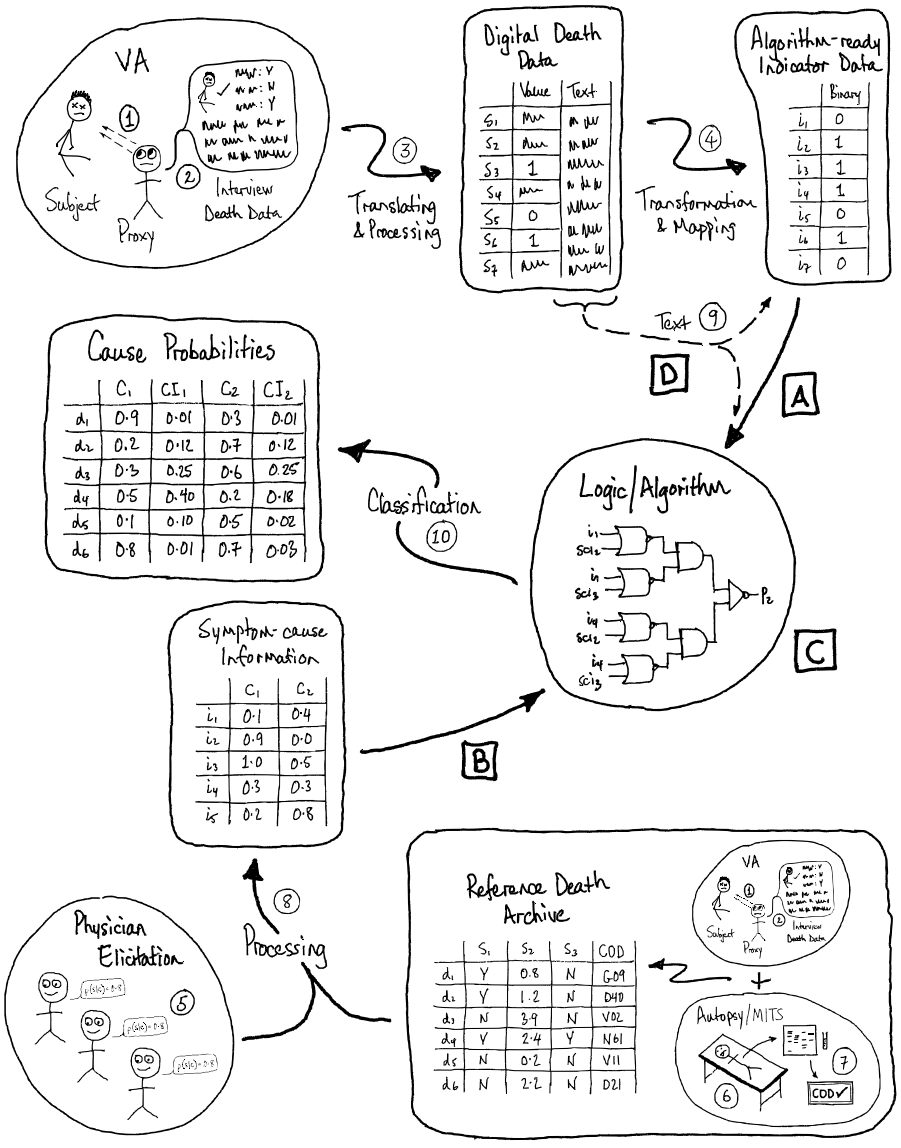
Burden of disease – BOD

- Overall indicator of population health
- Distribution of deaths by cause $\rightarrow$ prioritize causes
- Requires deaths to be registered and have a cause
- Much of Africa and parts of Asia and Latin America not able to run traditional vital statistics systems

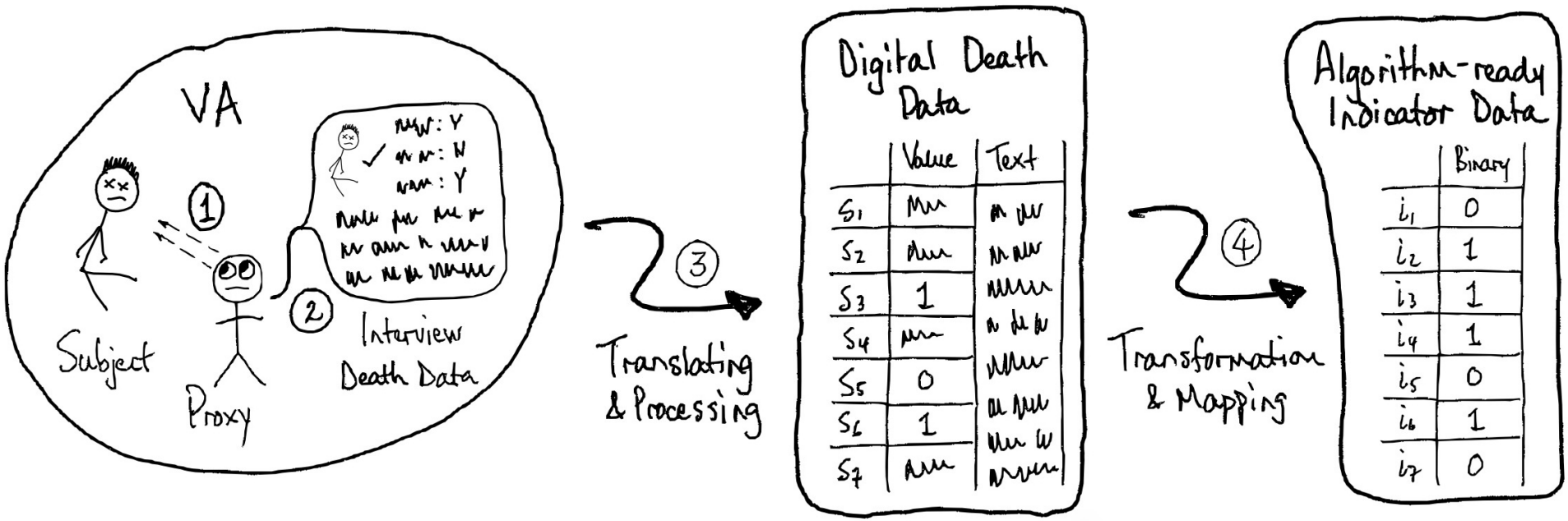
Verbal autopsy

- Feasible method to ascertain cause of death where traditional method are not possible
- With automated computer cause classification, cheap enough to be sustained in areas where it is needed

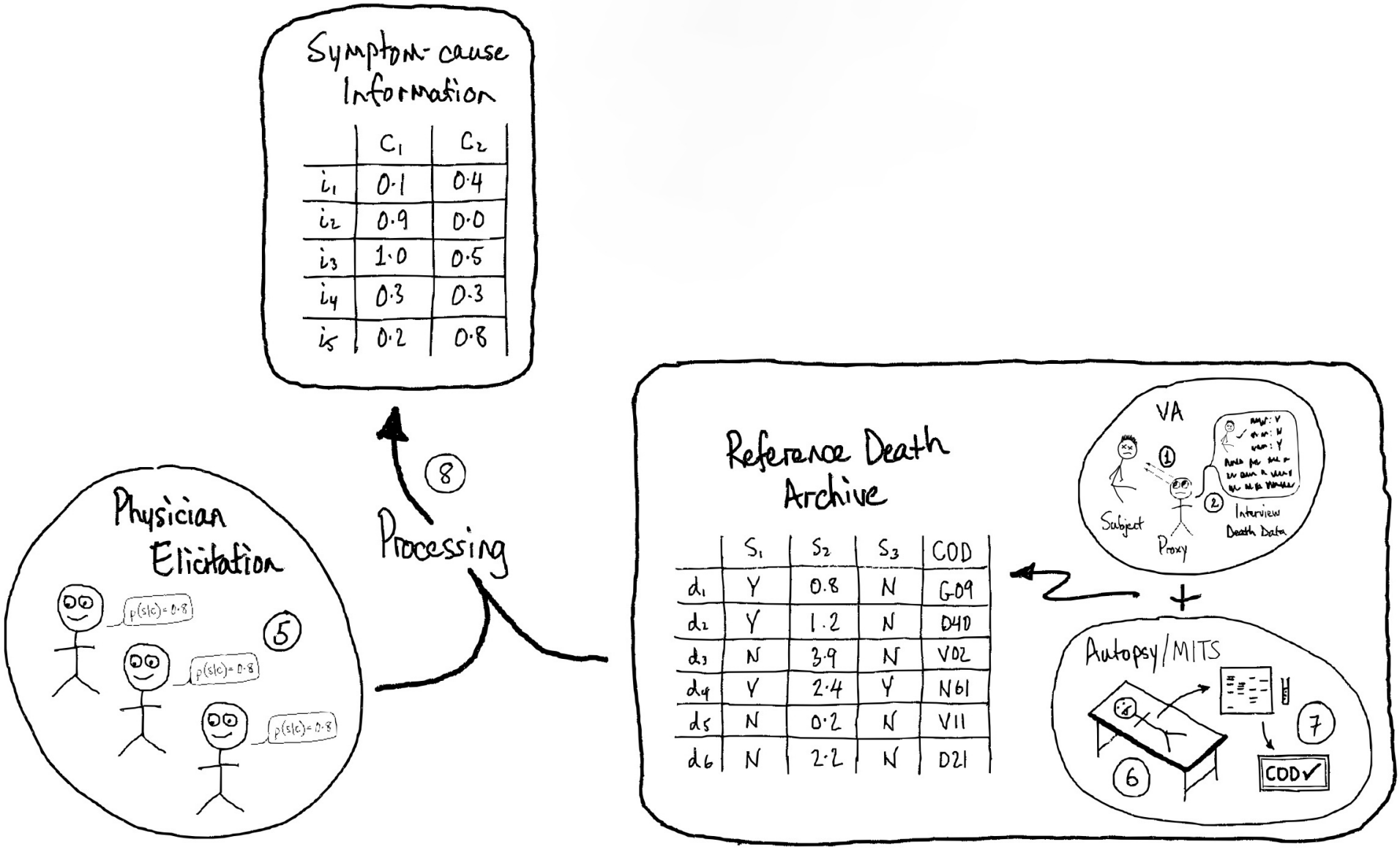
Verbal autopsy ecosystem



A: Verbal autopsy **interview and data**



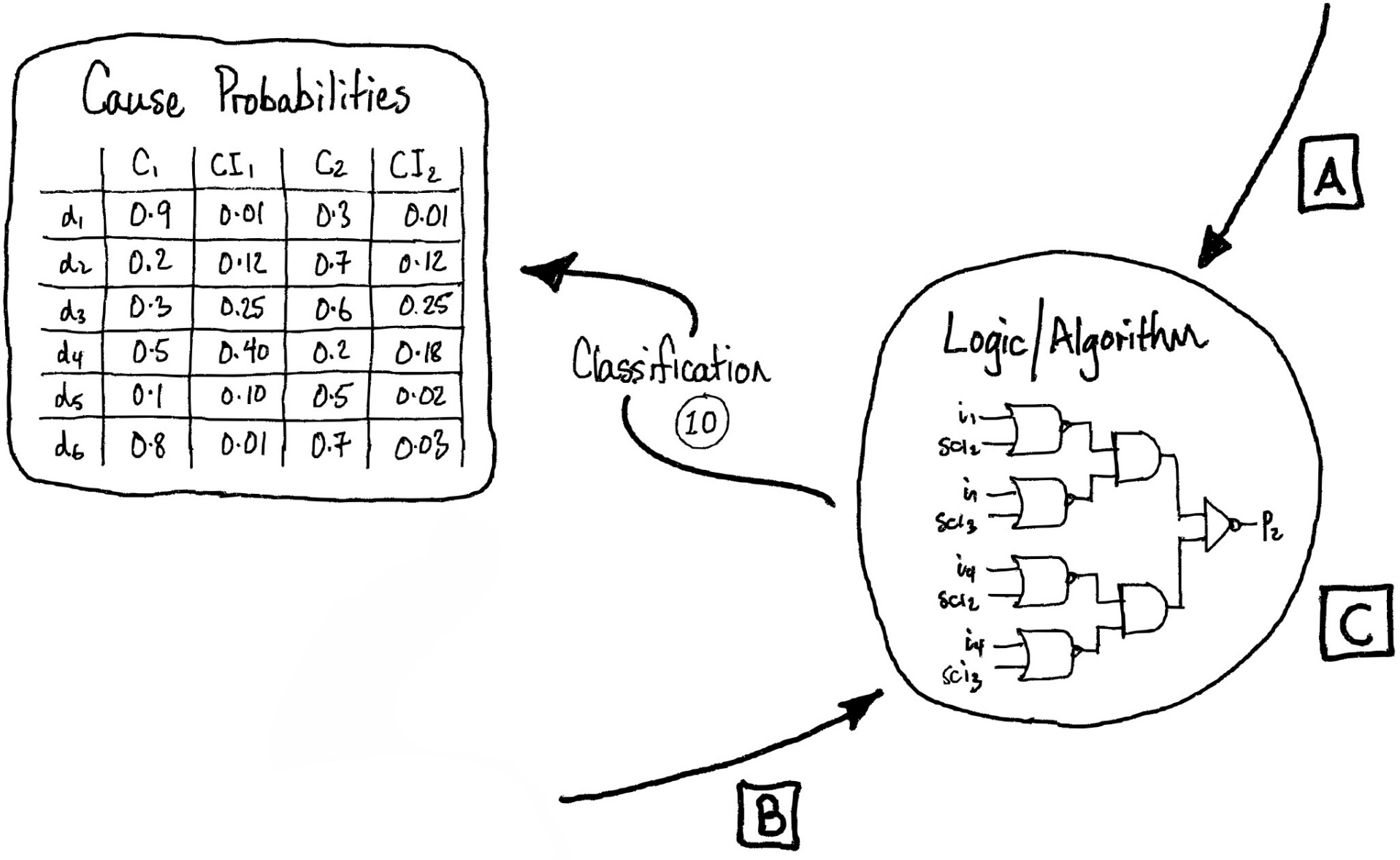
B: Verbal autopsy **symptom-cause information (SCI)**



	C_1	C_2
i_1	0·1	0·4
i_2	0·9	0·0
i_3	1·0	0·5
i_4	0·3	0·3
i_5	0·2	0·8

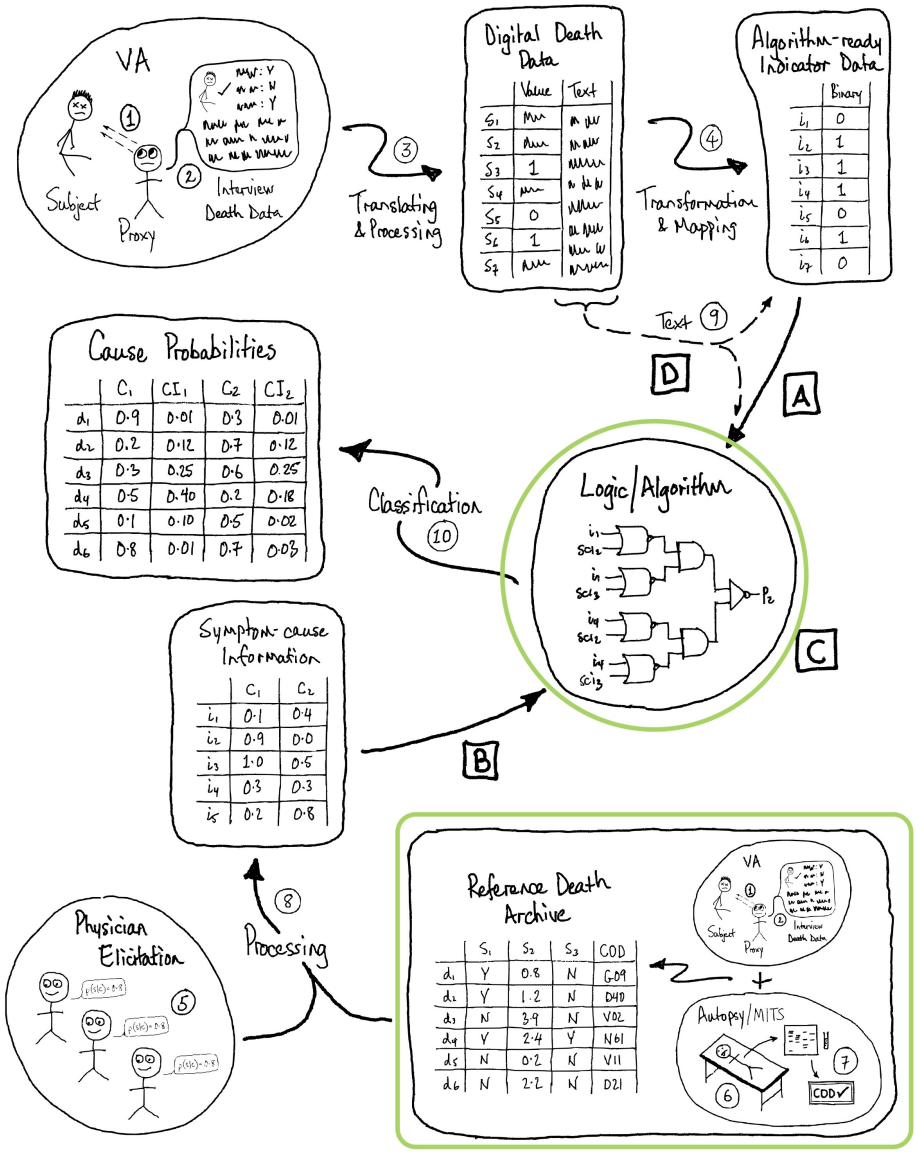
	S_1	S_2	S_3	COD
d_1	Y	0·8	N	G09
d_2	Y	1·2	N	D40
d_3	N	3·9	N	V02
d_4	Y	2·4	Y	N61
d_5	N	0·2	N	V11
d_6	N	2·2	N	D21

C: Verbal autopsy **cause-coding algorithms**



	C_1	CI_1	C_2	CI_2
d_1	0·9	0·01	0·3	0.01
d_2	0.2	0·12	0·7	0·12
d_3	0·3	0.25	0·6	0.25
d_4	0·5	0.40	0.2	0·18
d_5	0·1	0·10	0·5	0·02
d_6	0·8	0·01	0·7	0·03

Verbal autopsy ecosystem – focus on algorithms and SCI



Verbal autopsy summary

Aim: Assign a cause to a death with VA – classify the death using an abbreviated VA cause list

Data

1. Data from VA interview with knowledgeable caregiver of decedent
   - quantitative questions on signs, symptoms, diagnoses, durations, etc.
   - respondent's free-form narrative account of period leading up to death
2. Symptom-cause information (**SCI**) that describes the relationships between VA signs/symptoms and causes included in the VA cause list

Classification

1. Physicians review VA data and assign causes: **PCVA**
2. Automated statistical/computational algorithms assign causes using VA data *and* SCI: **CCVA**

VA is an imperfect and frustrating approach

Advantages

- FEASIBLE compared to traditional COD determination: autopsy, medical review, etc.
- Comparatively cheap
- Comparatively tractable – logistics, skills, etc.
- With computer coding:
  - does not require advanced skills
  - produces reproducible cause assignments in a timely fashion
  - no physician opportunity costs
- Capable of providing highly useful COD and BOD information for public health assessment and planning

Disadvantages

- Less accurate compared to traditional COD determination: autopsy, medical review, etc.
- Abbreviated cause list that does not easily mesh with full ICD cause lists, large catch-all causes
- Inherently low-information with many potential sources of error and bias: **classification is difficult**

VA Algorithms

VA cause-coding algorithms have three separable components

1. The VA data themselves
2. SCI that describes the relationship between VA symptoms and VA causes
3. The logic of the algorithm itself – mathematical, computational, statistical

The performance of each algorithm depends on both its logic and the SCI it uses

SCI can be swapped in/out and updated

This means that the performance of an algorithm can evolve and be adapted to a particular population

openVA Team contribution

Builds on InterVA

- InterVA is first widely used cause-coding algorithm for VA (Byass et al., 2019)
- Created and utilizes physician-elicited SCI as conditional probabilities: $\Pr(s|c)$
- Non-mathematical basis and only uses presence of a symptom
- Has unhelpful 'undetermined cause' – about 20% on average
- No uncertainty
- Incomplete and inaccurate description in literature and super clunky software

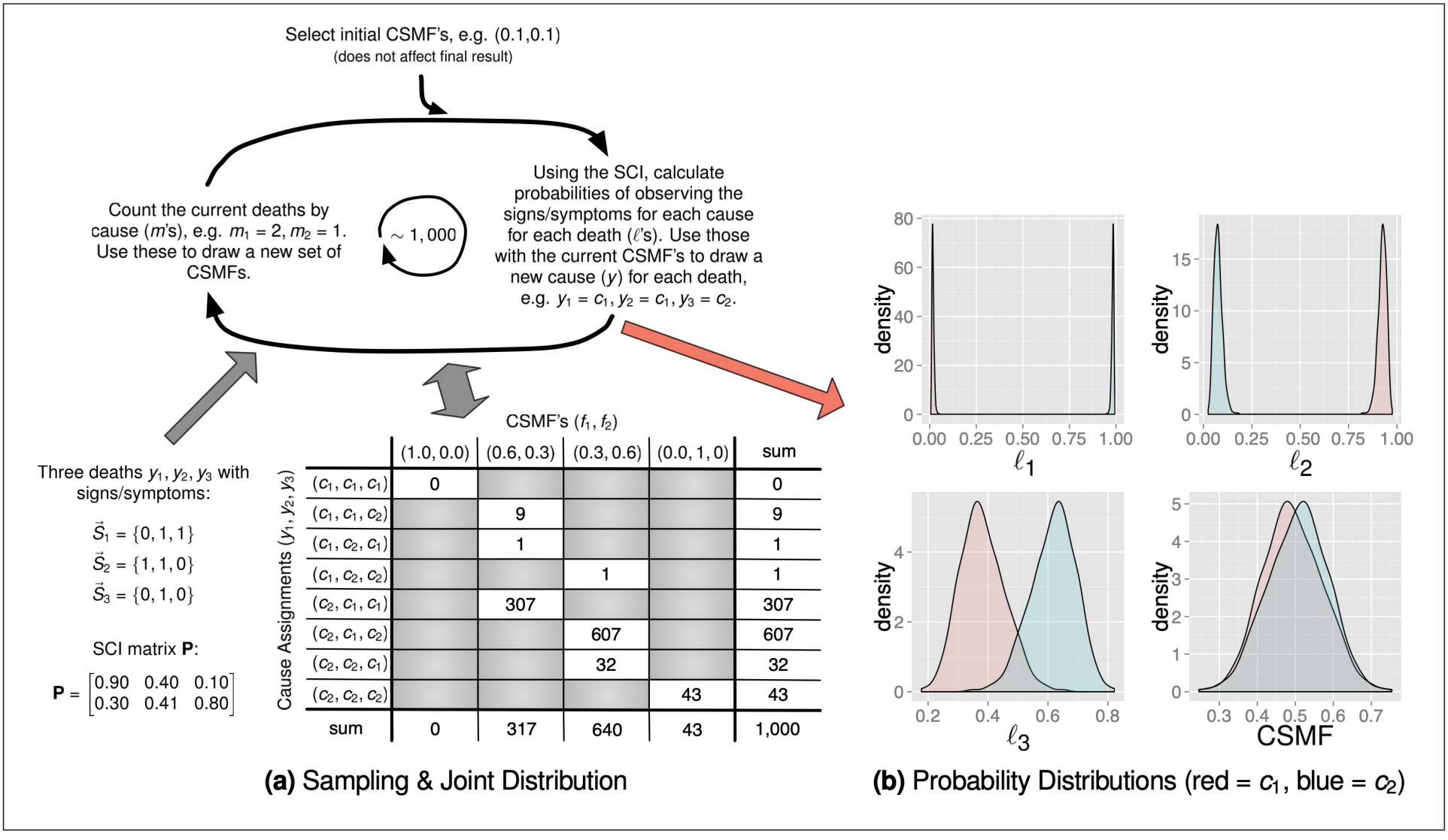
Our aim: create a new algorithm

- Principled mathematical/statistical basis $\rightarrow$ can trust it!
- Use both presence and absence of symptoms
- Estimate uncertainty/confidence for cause assignments and cause-specific mortality fractions
- No 'undetermined' causes – replace with uncertain/low confidence assignments

InSilicoVA

- With Tyler McCormick and Richard Li, I developed InSilicoVA (McCormick et al., 2016)
- Probabilistic model that estimates the joint distribution between individual-level cause classification and population cause-specific mortality fractions
- Uses InterVA's SCI
- Yields consistent probability distributions for all causes for each death **and** for each cause-specific mortality fraction
- No 'indeterminate cause' and uncertainty/confidence for everything

InSilicoVA heuristic sketch



	(1.0, 0.0)	(0.6, 0.3)	(0.3, 0.6)	(0.0, 1, 0)	sum
(c_1, c_1, c_1)	0				0
(c_1, c_1, c_2)		9			9
(c_1, c_2, c_1)		1			1
(c_1, c_2, c_2)			1		1
(c_2, c_1, c_1)		307			307
(c_2, c_1, c_2)			607		607
(c_2, c_2, c_1)			32		32
(c_2, c_2, c_2)				43	43
sum	0	317	640	43	1,000

(a) Sampling & Joint Distribution

(b) Probability Distributions (red = c_1, blue = c_2)

(a) Top: InSilicoVA sampling joint distribution of cause assignments and CSMFs . Bottom: data, SCI and 1,000 samples consistent with two causes of death $\{c_1, c_2\}$ and three deaths $\{y_1, y_2, y_3\}$. **(b) InSilicoVA output:** corresponding to the sample in (a), *estimated distributions* of individual probabilities ℓ of being assigned cause c_1 or c_2 and estimated distributions of the CSMFs for c_1 and c_2.

InSilicoVA main publication (McCormick et al., 2016)

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 515, 1036–1049, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1152191

Taylor & Francis
Taylor & Francis Group

Probabilistic Cause-of-Death Assignment Using Verbal Autopsies

Tyler H. McCormick[a,*], Zehang Richard Li[b,*], Clara Calvert[c], Amelia C. Crampin[c], Kathleen Kahn[d], and Samuel J. Clark[e,*]

[a]Department of Statistics and Sociology, University of Washington, Seattle, WA, USA; [b]Department of Statistics, University of Washington, Seattle, WA, USA; [c]London School of Hygiene and Tropical Medicine, London, UK; [d]MRC/Wits Rural Public Health and Health Transitions Unit, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa; [e]Department of Sociology, the Ohio State University, MRC/Wits Rural Public Health and Health Transitions Unit, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa, ALPHA Network, London School of Hygiene and Tropical Medicine, London, UK, and INDEPTH Network, Accra, Ghana

ABSTRACT

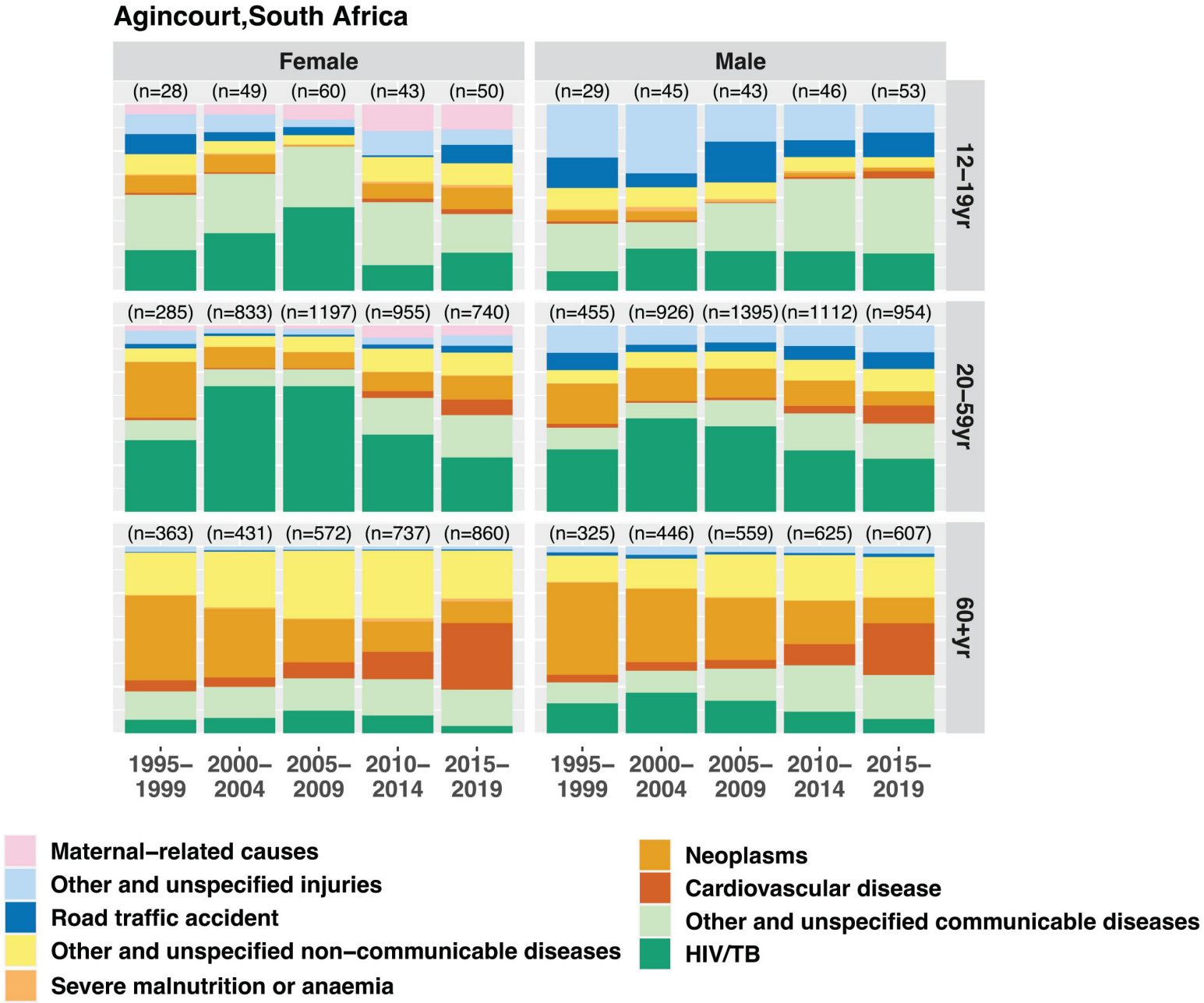
In regions without complete-coverage civil registration and vital statistics systems there is uncertainty about even the most basic demographic indicators. In such regions, the majority of deaths occur outside hospitals and are not recorded. Worldwide, fewer than one-third of deaths are assigned a cause, with the least information available from the most impoverished nations. In populations like this, verbal autopsy (VA) is a commonly used tool to assess cause of death and estimate cause-specific mortality rates and the distribution of deaths by cause. VA uses an interview with caregivers of the decedent to elicit data describing the signs and symptoms leading up to the death. This article develops a new statistical tool known as *InSilicoVA* to classify cause of death using information acquired through VA. InSilicoVA shares uncertainty between cause of death assignments for specific individuals and the distribution of deaths by cause across the population. Using side-by-side comparisons with both observed and simulated data, we demonstrate that InSilicoVA has distinct advantages compared to currently available methods. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received November 2014
Revised December 2015

KEYWORDS
Bayesian methods; cause of death; Demography; Verbal autopsy; Vital records

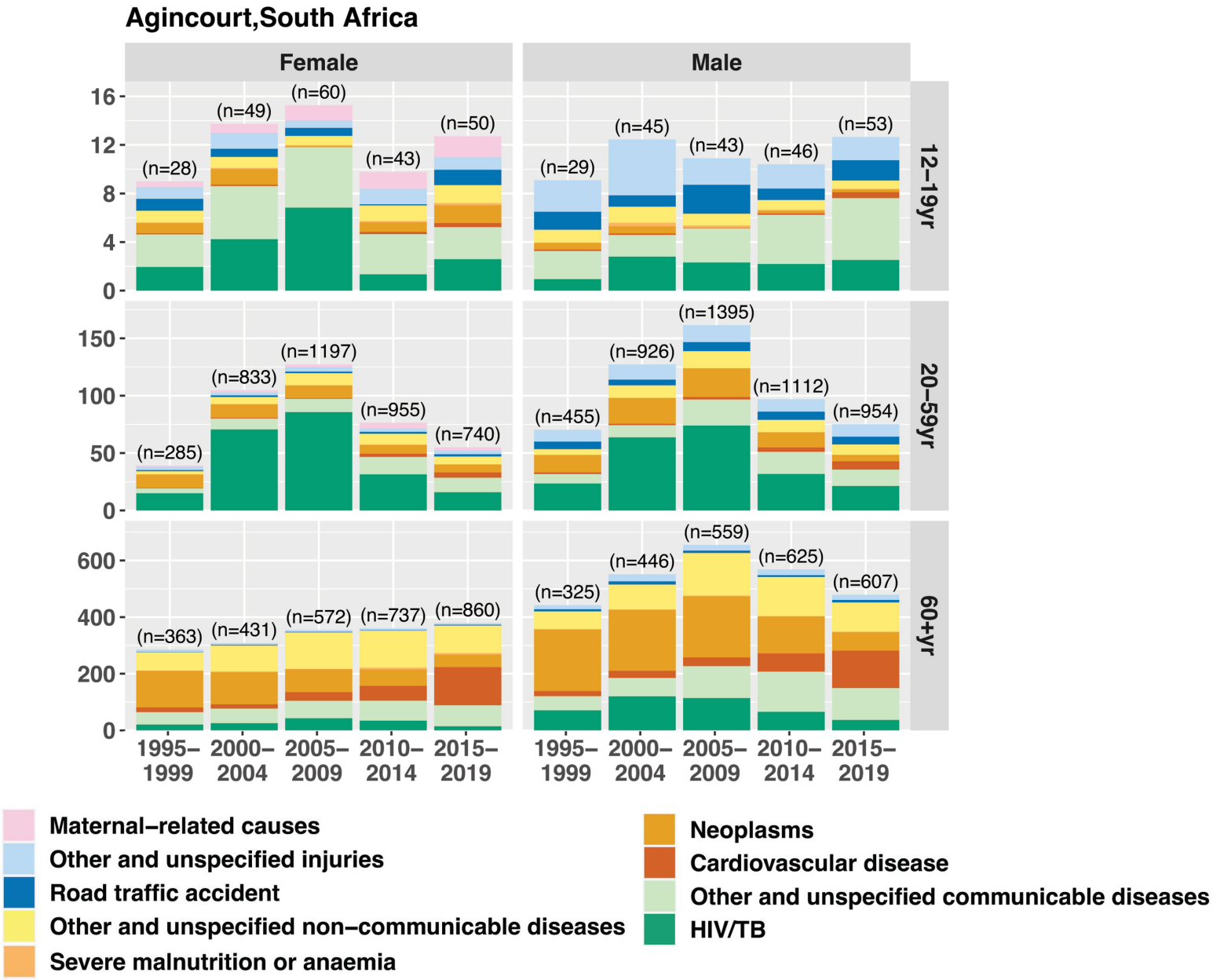
Example results: ALPHA Network COD distributions - Agincourt HDSS

(Chu et al., 2024)



Example results: ALPHA Network CASMR - Agincourt HDSS

(Chu et al., 2024)



Alpha Network COD analysis publication (Chu et al., 2024)

THE LANCET
Global Health

ARTICLES · Volume 12, Issue 8, E1278-E1287, August 2024 · ***Open Access*** Download Full Issue

Temporal changes in cause of death among adolescents and adults in six countries in eastern and southern Africa in 1995–2019: a multi-country surveillance study of verbal autopsy data

Yue Chu, MSPH [a,b,c] · Milly Marston, PhD [d] · Albert Dube, MSc [e] · Charles Festo, MSc [f] · Eveline Geubbels, PhD [f] · Prof Simon Gregson, DPhil [g,h] · Kobus Herbst, MSc [i,j] · Chodziwadziwa Kabudula, PhD [k] · Prof Kathleen Kahn, MD [k] · Tom Lutalo, PhD [l] · Louisa Moorhouse, PhD [g] · Prof Robert Newton, PhD [m,n] · Constance Nyamukapa, PhD [g,h] · Ronald Makanga, PhD [m] · Emma Slaymaker, PhD [d] · Mark Urassa, MSc [o] · Abdhalah Ziraba, PhD [p] · Clara Calvert, PhD [d,q,†] · Prof Samuel J Clark, PhD [a,b,c,k,†]
Show less

Affiliations & Notes · Article Info · Linked Articles (1)

Download PDF · Cite · Share · Set Alert · Get Rights · Reprints

Software: the openVA Suite

The openVA Team has developed and supports a range of software for VA, including InSilicoVA

- openVA: https://cran.r-project.org/package=openVA
- InSilicoVA: https://cran.r-project.org/package=InSilicoVA
- interVA5: https://cran.r-project.org/package=InterVA5
- interVA4: https://cran.r-project.org/package=InterVA4
- Tariff 1: https://cran.r-project.org/package=Tariff
- CrossVA: https://cran.r-project.org/package=CrossVA
- pyCrossVA: https://pypi.org/project/pycrossva/0.92/
- pyOpenVA: https://github.com/verbal-autopsy-software/pyopenva_GUI
- openVA Pipeline: https://pypi.org/project/openva-pipeline/
- Others: https://github.com/verbal-autopsy-software
- User-oriented description and tutorial – **The openVA Toolkit for Verbal Autopsies** (Li et al., 2023)

The openVA Suite is the reference implementation of VA algorithms that support WHO VA standards and is used by a wide variety of researchers and CRVS organizations globally

openVA Toolkit publication (Li et al., 2023)

The R Journal

HOME ISSUES NEWS CURRENT SUBMIT

The openVA Toolkit for Verbal Autopsies

Abstract:

Verbal autopsy (VA) is a survey-based tool widely used to infer cause of death (COD) in regions without complete-coverage civil registration and vital statistics systems. In such settings, many deaths happen outside of medical facilities and are not officially documented by a medical professional. VA surveys, consisting of signs and symptoms reported by a person close to the decedent, are used to infer the COD for an individual, and to estimate and monitor the COD distribution in the population. Several classification algorithms have been developed and widely used to assign causes of death using VA data. However, the incompatibility between different idiosyncratic model implementations and required data structure makes it difficult to systematically apply and compare different methods. The openVA package provides the first standardized framework for analyzing VA data that is compatible with all openly available methods and data structure. It provides an open-source, R implementation of several most widely used VA methods. It supports different data input and output formats, and customizable information about the associations between causes and symptoms. The paper discusses the relevant algorithms, their implementations in R packages under the openVA suite, and demonstrates the pipeline of model fitting, summary, comparison, and visualization in the R environment.

cite pdf tweet

What we are doing now

VA Algorithms

- Support dependence among symptoms
- Create domain-adaptive algorithm
- Automated text analysis integrated into algorithm
- Improve SCI with information on dependence and potentially other informative covariates

Software - responding to user feedback

- Need **much** easier to use software
  - Real world users cannot use `R`, Java, C, C++, Python, etc. – keeping all the tools updated and integrated is way too much
  - Must create pre-compiled, user friendly software with carefully thought through point-and-click graphical interfaces – pyOpenVA
  - Must seamlessly integrate software into existing work/data flows
- Integrate into dashboards and existing data streams and data stores – openVA pipeline

Completely new context and application – HCAP

Harmonized cognitive assessment protocol or (HCAP)

- HCAP very similar to VA – quantitative and free-form information from proxy respondent used to classify individual into categories of cognitive function
- HCAP used in all settings, high income and LMIC
- High and increasing demand to rapidly (re)assess cognitive function as people age
- Current system like VA 20 years ago – unreliable and very labor intensive

New approaches and text analysis

Text analysis in algorithm

- NLP/LLM methods to classify causes using text from narrative account
- NLP/LLM methods to produce additional indicators for algorithms, similar to existing indicators from quantitative symptoms
- Improve narrative account in interviews to work better with automated methods
- Adapt or create new NLP/LLM methods specifically for VA

LLM/Chatbots

- Develop chatbot-driven semi-structured interview, maybe fully automated
- Or, restrict chatbot input to elicitation of narrative account
- Develop semi-automated, realtime classification interview using different/bespoke structured approach, stop interview when classification is precise/confident enough

All of these approaches require training data, both general and VA-specific

New and more informative supporting data

Machine learning approaches need training data, and existing/new algorithms need SCI that represents dependent relationships between symptoms and causes

Existing SCI is elicited from physicians as $\Pr(s|c)$

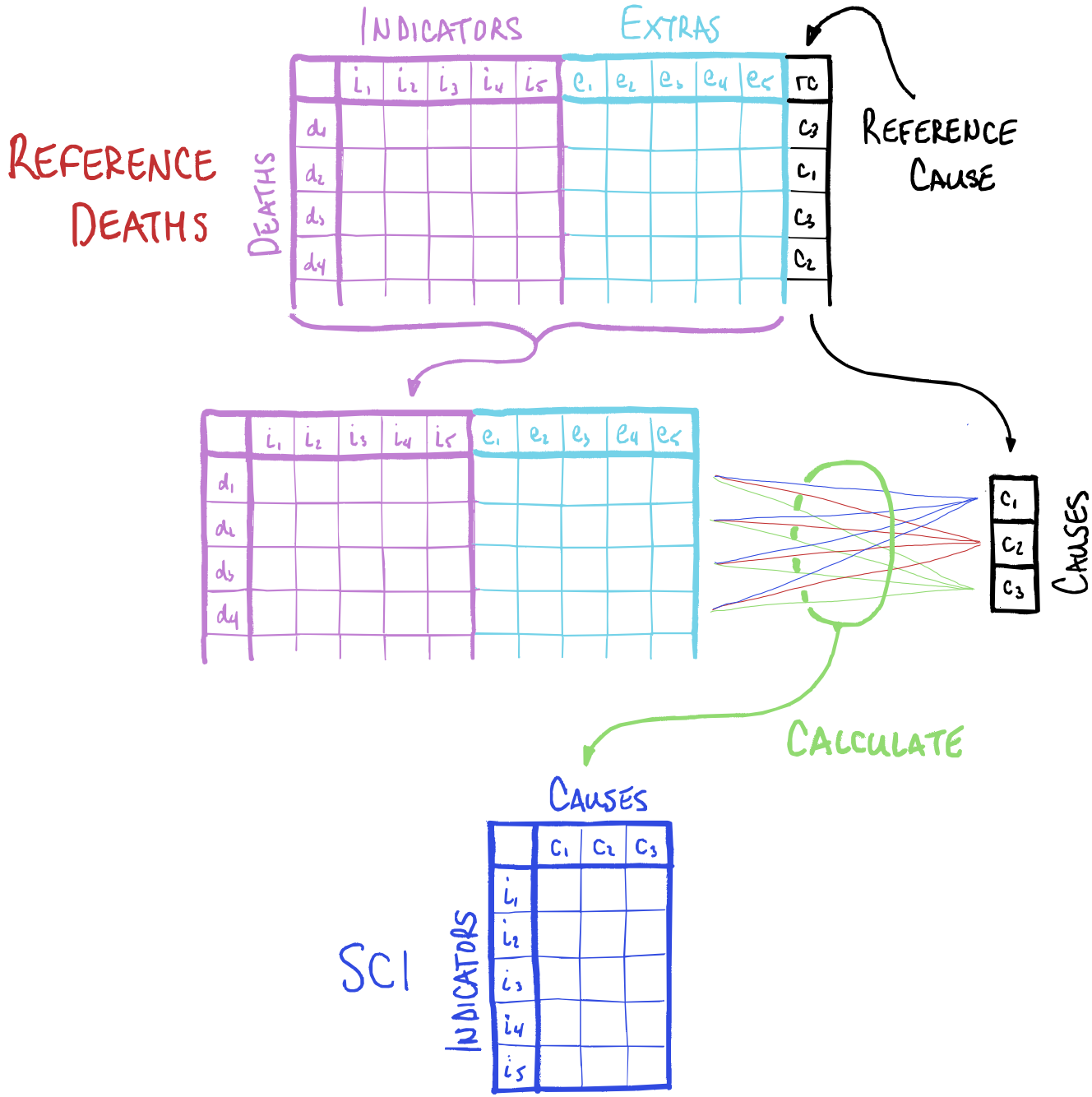
An alternative is a collection reference deaths having both a standard VA and a reference cause, and potentially having other predictive covariates

Reference deaths

New project: the Reference Death Archive (RDA)

- Hosted by WHO in Geneva and Africa Health Research Institute (**AHRI**), Somkhele and Durban, South Africa
- Deaths with VA and reliable cause from other cause attribution method from variety of research projects and Brazilian mortality surveillance system in state of Sao Paulo
- Many deaths include information from minimally-invasive tissue sample (**MITS**) as informative covariates
- Deaths from mortality surveillance system in Sao Paulo Brazil (reference cause is autopsy), CHAMPS, COMSA, and MITS Alliance sites – all include MITS
- Majority of deaths from Brazil where systems cover very large populations
- Within about five years, sufficient deaths to support machine learning approaches and create much better SCI

SCI from reference deaths and covariates



RDA infrastructure summary – at WHO headquarters

Trusted researcher environment

- Contains cleaned, harmonized, documented microdata with instruments and ethical clearance documents
- De-identified microdata come in but **cannot ever leave in raw, individual form**
- Free access to all microdata and documentation within secure environment
- JupyterHub: Julia, R, Python and potentially more
- On demand high-performance computing for Bayesian methods, machine learning, or AI
- **Aggregated** data products can be submitted for inspection and sharing in public-facing NADA repository

Separate NADA data catalog and repository of data sets

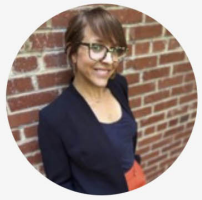
- Receives certified aggregated data products from JupyterHub – must be fully documented
- Manages sharing aggregated data products
- Browse/searchable metadata and data catalogs
- Handles sharing transactions, e.g. licensing and data use agreements

openVA Team

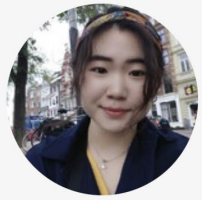
Research Team

Sam Clark

Principal Investigator



Nicole Angotti

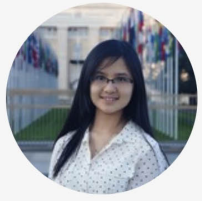


Yoonyoung Choi

Collins Ochieng

Isaac Lyatuu

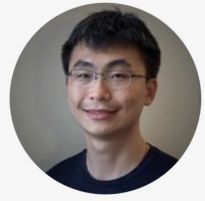
Sherry Zhao



Yue Chu

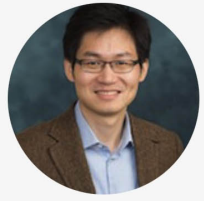
Jason Thomas

Tyler McCormick



Zehang Richard Li

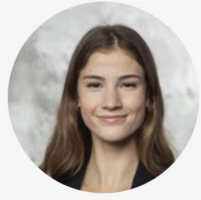
Clarissa Surek-Clark



Zhenke Wu

Emeritus Members

Eungang Peter Choi



Melina Raglin

Contacts and information

Contacts

- work@samclark.net
- https://samclark.net
- http://openva.net
- https://github.com/verbal-autopsy-software
- info@openva.net
- help@openva.net
- These slides: https://samclark.net/ut-austin2025

Funders

- National Institute for Child Health and Human Development (NICHD), USA NIH
- Bill and Melinda Gates Foundation
- USA CDC – International
- Vital Strategies
- OSU Institute for Population Research (IPR)
- UN Population Division (UNPD)

References I

Byass, P., L. Hussain-Alkhateeb, L. D'Ambruoso, S. Clark, J. Davies, E. Fottrell, J. Bird, C. Kabudula, S. Tollman, K. Kahn, L. Schiöler, and M. Petzold (2019, May). An integrated approach to processing WHO-2016 verbal autopsy data: the InterVA-5 model. *BMC medicine 17*(1), 102.

Chu, Y., M. Marston, A. Dube, C. Festo, E. Geubbels, S. Gregson, K. Herbst, C. Kabudula, K. Kahn, T. Lutalo, L. Moorhouse, R. Newton, C. Nyamukapa, R. Makenga, E. Slaymaker, M. Urassa, A. Ziraba, C. Calvert, and S. J. CLARK† (2024). **Temporal Changes in Cause of Death among Adolescents and Adults in Six Countries in Eastern and Southern Africa: a Multi-country Cohort Study using Verbal Autopsy Data**. *Lancet Global Health 12*(e1278 – 87).

Golub, G. H., A. Hoffman, and G. W. Stewart (1987). A generalization of the eckart-young-mirsky matrix approximation theorem. *Linear Algebra and Its Applications 88*, 317–327.

Good, I. J. (1969). Some applications of the singular decomposition of a matrix. *Technometrics 11*(4), 823–831.

Li, Z. R., J. Thomas, E. Choi, T. H. McCormick, and S. J. CLARK† (2023). **The openVA Toolkit for Verbal Autopsies**. *The R Journal*.

Li, Z. R., J. Thomas, E. Choi, T. H. McCormick, and S. J. Clark (2023). The openVA Toolkit for Verbal Autopsies. *The R Journal*.

McCormick, T. H., Z. R. Li, C. Calvert, A. C. Crampin, K. Kahn, and S. J. CLARK† (2016). **Probabilistic Cause-of-Death Assignment using Verbal Autopsies**. *Journal of the American Statistical Association 111*(515), 1036 – 1049.

Nichols, E. K., P. Byass, D. Chandramohan†, S. J. CLARK, A. D. Flaxman, R. Jakob, J. Leitao, N. Maire, C. Rao, I. Riley, and P. Setel (2018). **The WHO 2016 Verbal Autopsy Instrument: An International Standard Suitable for Automated Analysis by InterVA, InsilicoVA, and Tariff 2.0**. *PLoS Medicine 15*(1), e1002486.

References II

Stewart, G. W. (1993). On the early history of the singular value decomposition. *SIAM review 35*(4), 551–566.

Strang, G. (2009). *Introduction to Linear Algebra 4e*. Wellesley-Cambridge Press.