Data Science and Math in Global Health: Two Examples

Samuel J. Clark

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

MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health,
Faculty of Health Sciences, University of the Witwatersrand

BIOMATH 2024

June 20, 2024
Scottburgh, South Africa

Overview

Introduction

Example 1: Age-specific Mortality Models

Example 2: Verbal Autopsy

Closing

About me

Not a mathematician! Demographer → Epidemiologist, Data Scientist/Statistician. Most of my career working on issues affecting Africa.

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

Overall objective

Improve global health

What is global health

- Population-level
- Description of health, epidemiology, and demography of populations
- Create and monitor interventions to improve specific aspects of individual and population health

What we want to do

- Describe: measure, quantify
- Understand: analysis, experiments
- Inform: communicate to decision makers

The data science life cycle is a nice way to think about these activities together

Affect the world with research, models, and software

Data science life cycle

- Framework for applying and improving data-driven methods to affect the world
- Potentially includes all four general types of academic pursuit
  - Descriptive: thoroughly document the real, uncontrolled world, produce structured data
  - Theoretical: explore concepts and create new ideas, often using data from descriptive studies
  - Experimental: conduct structured, controlled studies to investigate the world, often using theory and results from descriptive studies
  - Application/engineering: apply knowledge and theory to reliably affect the world
- Fundamentally a loop – sequence of activities that cycles to produce and improve the desired effects
- Useful because it expands and organizes standard concept of academic work and provides clear interfaces to non-academic collaborators
- **Outcomes are more effective when all these steps are integrated and specialists in each talk to each other**

The data science life cycle

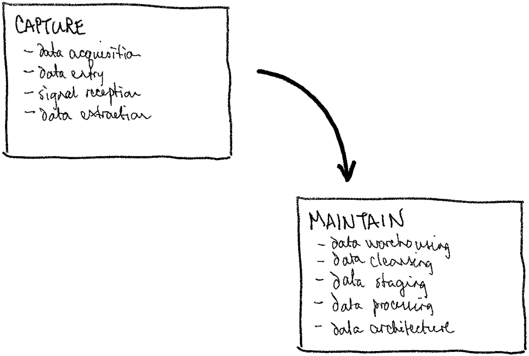
(Ref: Berkeley School of Information: What is Data Science?)

CAPTURE

- data acquisition
- data entry
- signal reception
- data extraction

The data science life cycle

(Ref: Berkeley School of Information: What is Data Science?)

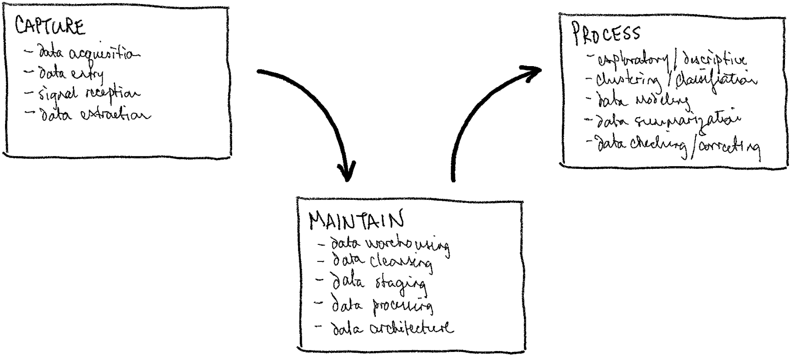


CAPTURE
- data acquisition
- data entry
- signal reception
- data extraction

MAINTAIN
- data warehousing
- data cleansing
- data staging
- data processing
- data architecture

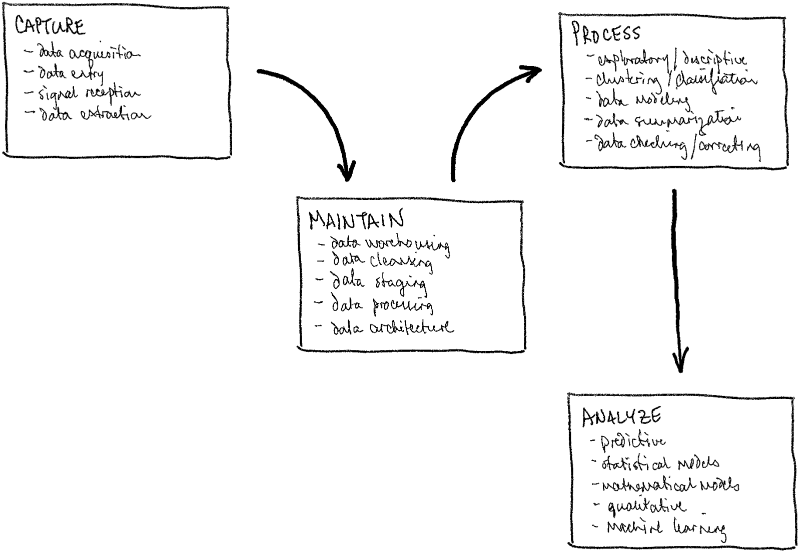
The data science life cycle

(Ref: Berkeley School of Information: What is Data Science?)



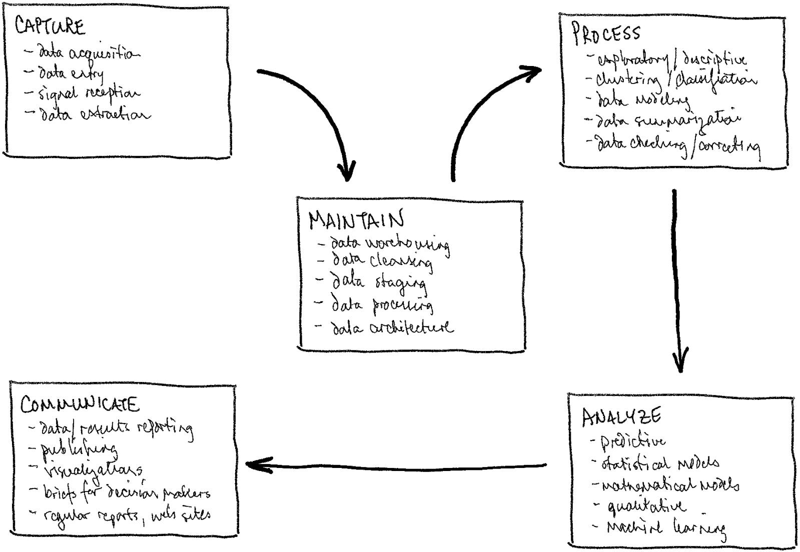
The data science life cycle

(Ref: Berkeley School of Information: What is Data Science?)



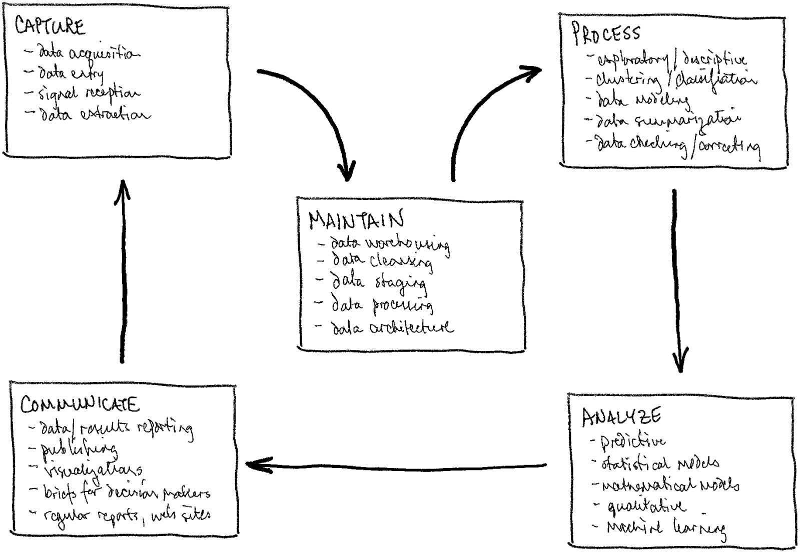
The data science life cycle

(Ref: Berkeley School of Information: What is Data Science?)



The data science life cycle

(Ref: Berkeley School of Information: What is Data Science?)



CAPTURE
- data acquisition
- data entry
- signal reception
- data extraction

PROCESS
- exploratory / descriptive
- clustering / classification
- data modeling
- data summarization
- data checking / correcting

MAINTAIN
- data warehousing
- data cleansing
- data staging
- data processing
- data architecture

ANALYZE
- predictive
- statistical models
- mathematical models
- qualitative
- machine learning

COMMUNICATE
- data/results reporting
- publishing
- visualizations
- briefs for decision makers
- regular reports, web sites

This talk

Examples of math/computational methods developed as part of global health data science life cycle

1. Mathematical/statistical model of age-specific mortality
   - Audience: UN Population Division and bi-annual World Population Prospects report of global population estimates and forecasts
2. Verbal autopsy cause of death ascertainment for civil registration and vital statistics
   - Audience: epidemiologists, national/local governments, World Health Organization (**WHO**)

Introduction

Example 1: Age-specific Mortality Models

Example 2: Verbal Autopsy

Closing

Mortality

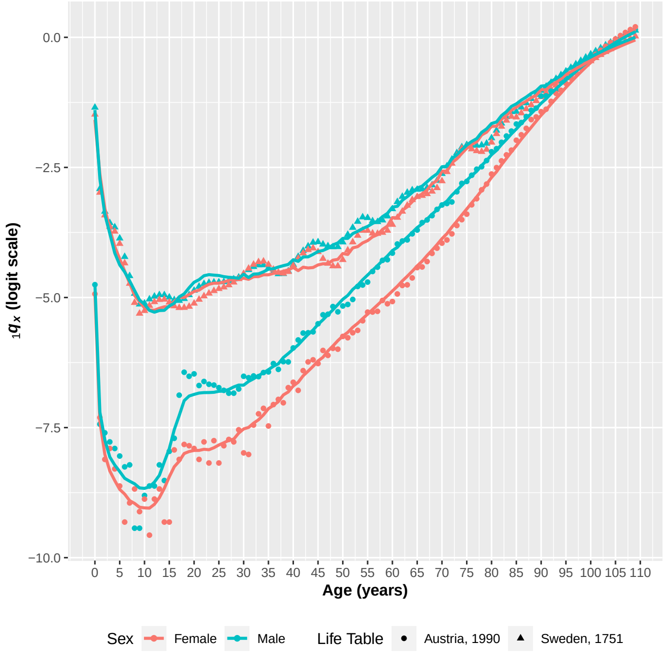
Age-specific mortality

- 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 and 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 age schedules of mortality



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 $\mathbf{X}$ (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): ∼ 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

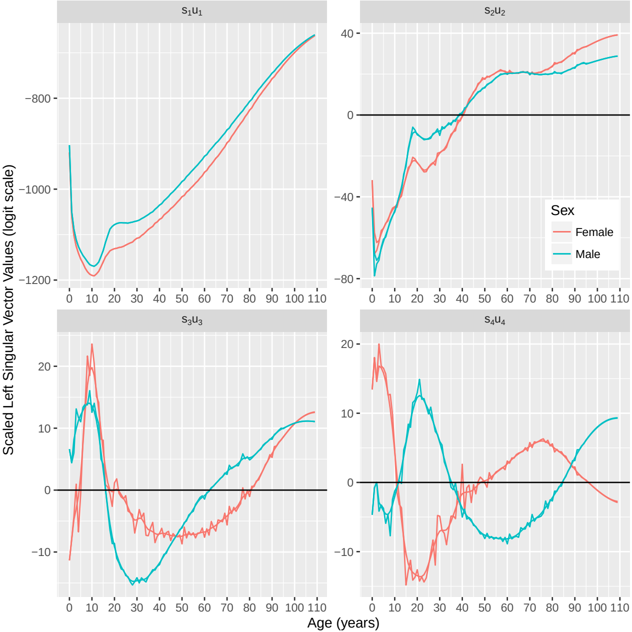
- SVD of HMD yields
  - Constant age-varying components - LSVs that represent systematic shape of human mortality by age
  - Varying weights - RSVs whose elements are the weights necessary in Equation 2 to fully reconstruct the original HMD mortality schedules

Prediction

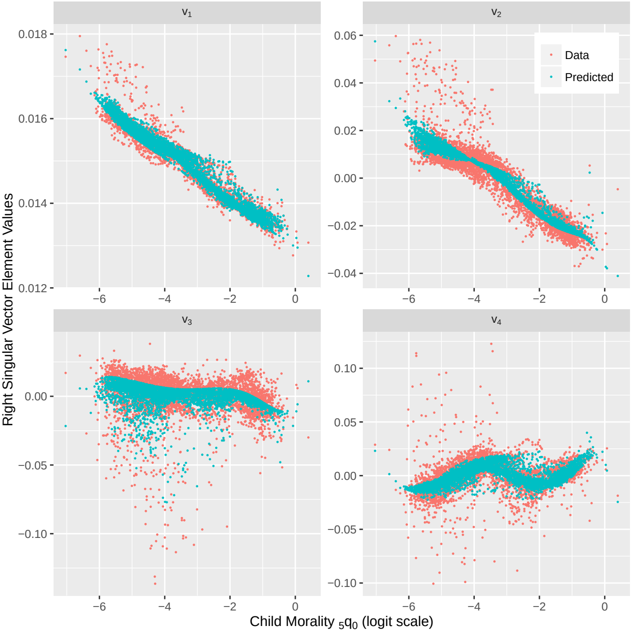
- Build statistical models that relate weights (RSV elements) to interesting predictor that varies with human mortality
- Use the statistical models to predict weights as functions of the predictor, and use the weights to reconstruct full schedules of age-specific mortality

Super parsimonious, efficient model of 111 1-year age-specific mortality values

LSVs from SVD of HMD



Relationship between RSV element weights and level of child mortality



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

Mortality model data science life cycle

Stages of the data science life cycle

- Need defined by UN Population Division and variety of research in epidemiological and demographic literature
- Data collection: HMD and UN Population Division
- Data management, cleaning, validating etc: HMD
- Data processing, summarizing: our team
- Data analysis: our team developed new methods/model
- Communication: journal publication, all code used for paper, `R` package for final model, presentations to UN Population Division – **all are critical**
- *Looping again:* talking with UN Population Division and refining data, model, and software . . .

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`
- `R` package

Introduction

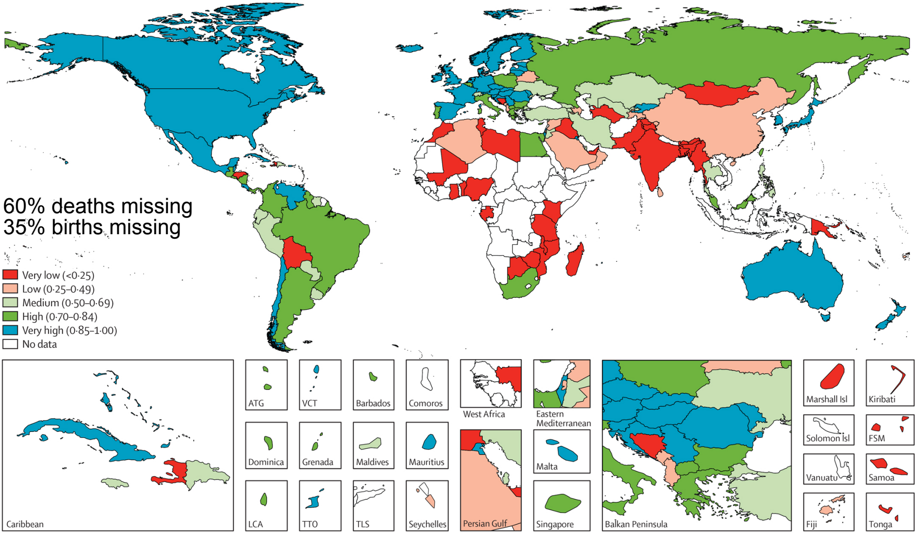
Example 1: Age-specific Mortality Models

Example 2: Verbal Autopsy

Closing

Global civil registration and vital statistics – **burden of disease**

(Mikkelsen et al., 2015)



60% deaths missing
35% births missing

Verbal autopsy – **VA**

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

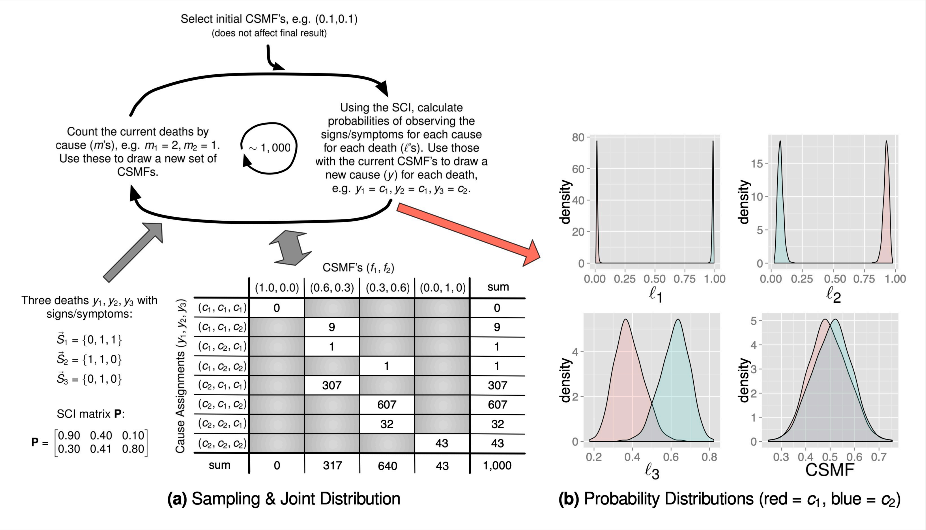
- Principled mathematical/statistical basis → can trust it!
- Use both presence and absence of symptoms
- Estimate uncertainty/confidence for cause assignments and cause-specific mortality fractions
- No 'undetermined' causes – replace with uncertain/low confidence assignments

InSilicoVA

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



Select initial CSMF's, e.g. (0.1,0.1)
(does not affect final result)

Using the SCI, calculate probabilities of observing the signs/symptoms for each cause for each death (ℓ's). Use those with the current CSMF's to draw a new cause (y) for each death, e.g. $y_1 = c_1, y_2 = c_1, y_3 = c_2$.

$\sim 1,000$

Count the current deaths by cause (m's), e.g. $m_1 = 2, m_2 = 1$. Use these to draw a new set of CSMFs.

Three deaths y_1, y_2, y_3 with signs/symptoms:

$\vec{S}_1 = \{0, 1, 1\}$

$\vec{S}_2 = \{1, 1, 0\}$

$\vec{S}_3 = \{0, 1, 0\}$

SCI matrix **P**:

$$\mathbf{P} = \begin{bmatrix} 0.90 & 0.40 & 0.10 \\ 0.30 & 0.41 & 0.80 \end{bmatrix}$$

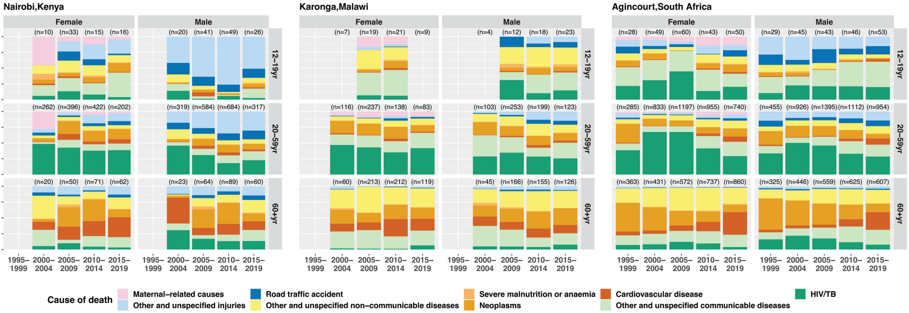
Cause Assignments (y_1, y_2, y_3)	CSMF's (f_1, f_2) $(1.0, 0.0)$	$(0.6, 0.3)$	$(0.3, 0.6)$	$(0.0, 1, 0)$	sum
(c_1, c_1, c_1)	0				0
(c_1, c_1, c_2)		9			9
(c_1, c_2, c_1)		1			1
(c_1, c_2, c_2)			1		1
(c_2, c_1, c_1)		307			307
(c_2, c_1, c_2)			607		607
(c_2, c_2, c_1)			32		32
(c_2, c_2, c_2)				43	43
sum	0	317	640	43	1,000

(a) Sampling & Joint Distribution

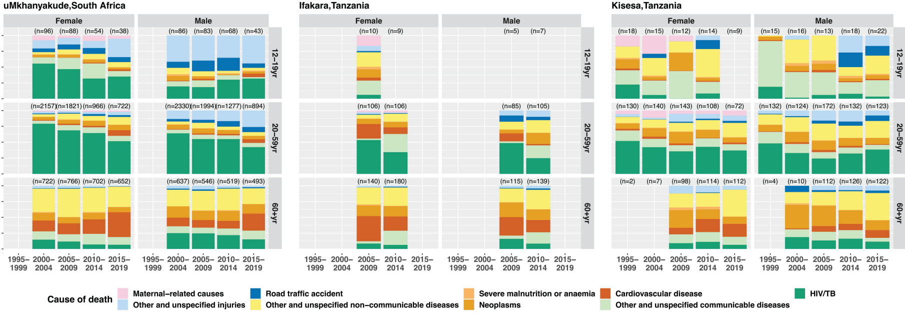
(b) Probability Distributions (red = c_1, blue = c_2)

(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.

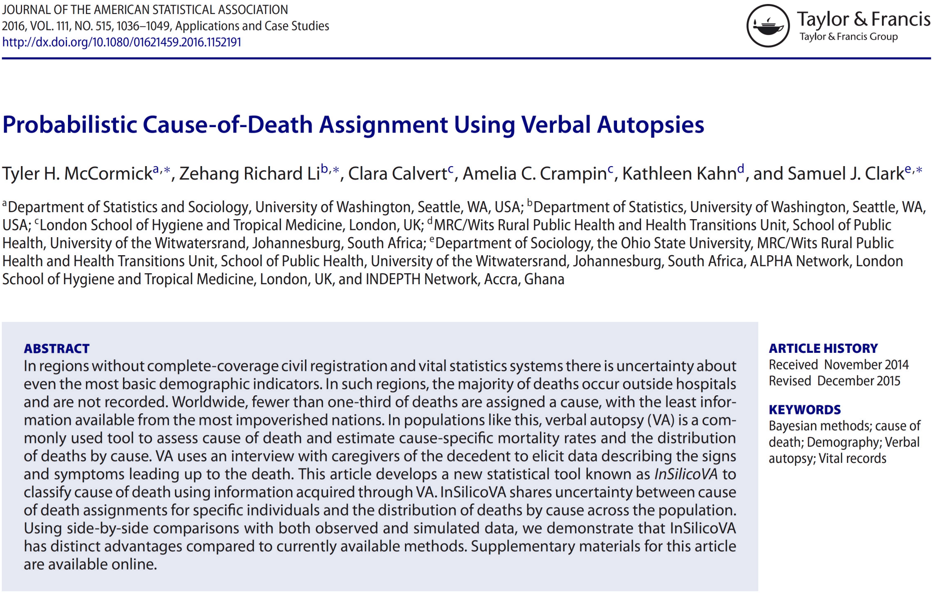
Example results: ALPHA Network cause-specific mortality fractions 1



Example results: ALPHA Network cause-specific mortality fractions 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

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

VA algorithms and data science life cycle

Stages data science life cycle

- Need defined by researchers/literature, WHO, and Data for Health Initiative (**D4H**) Partners with national governments
- Data collection: research projects, governments, openVA Team contributes to and supports global VA standards development at WHO
- Data management, cleaning, validating etc: research projects, governments, openVA Team
- Data processing, summarizing: research projects, governments, openVA Team
- Data analysis: openVA Team developed new methods/software
- Communication: journal publications, all code used for papers, `R` packages/Python modules/compiled applications for final methods, user tutorial publications and workshops, presentations to WHO and others – **all are critical**
- *Looping again:* talking with WHO, D4H, researchers, and governments to understand needs and reactions to current tools to refine data, methods, and software . . .

What are we doing now

Our ideas

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

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

Ongoing algorithm development priorities

Completely new things

- Adapt this approach to other problems, e.g. the '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 account
- NLP/LLM methods to produce additional indicators for algorithms, similar to existing indicators from quantitative symptoms
- Improve 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
- 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 $\Pr(s|c)$

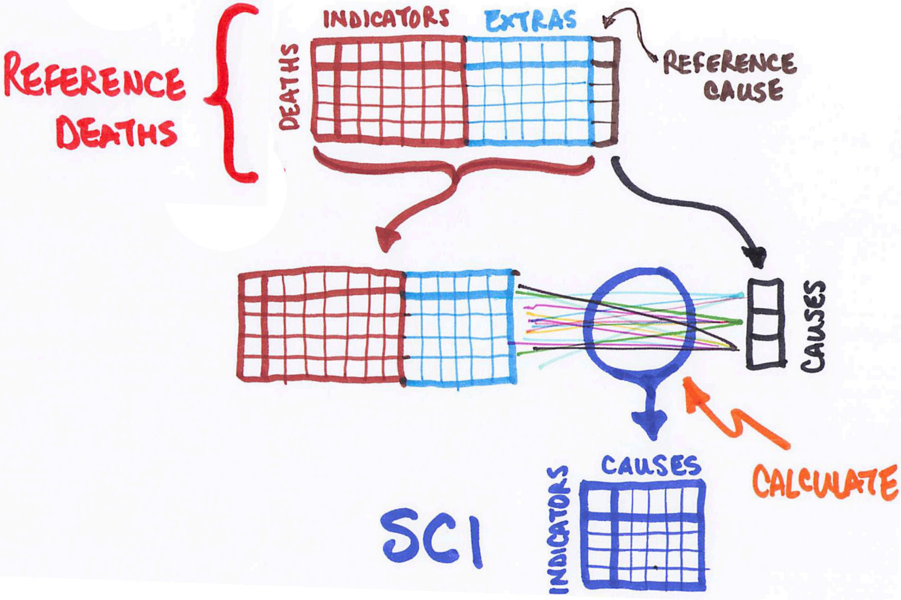
An alternative is a collection reference deaths having both a standard VA and a reference cause, and potentially having other predictive covariates

Reference deaths

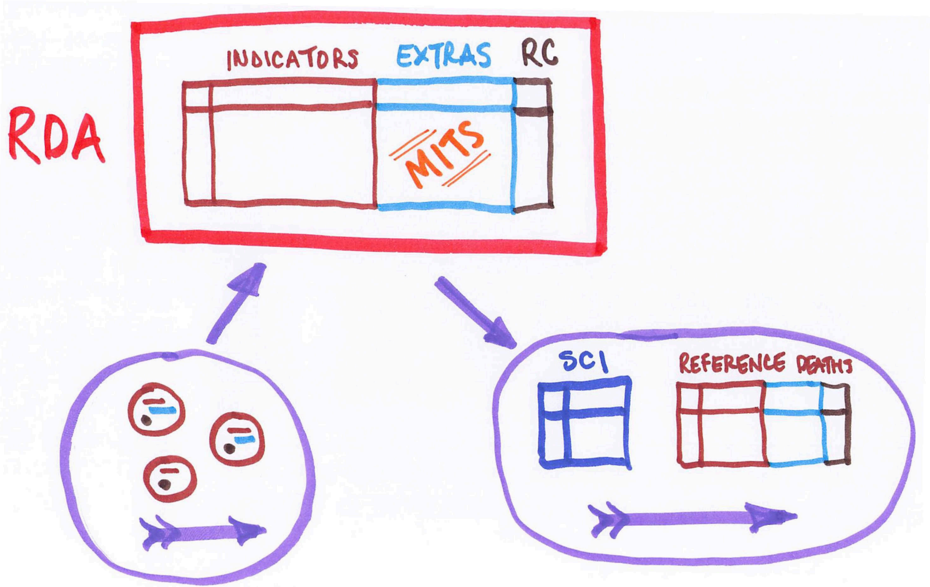
New project: the reference death archive (RDA)

- Hosted by WHO in Geneva and Africa Health Research Institute (**AHRI**), Somkhele and Durban, South Africa (just up the coast from Sottburgh)
- 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 Brazil include full traditional autopsy and MITS
- Majority of deaths form 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



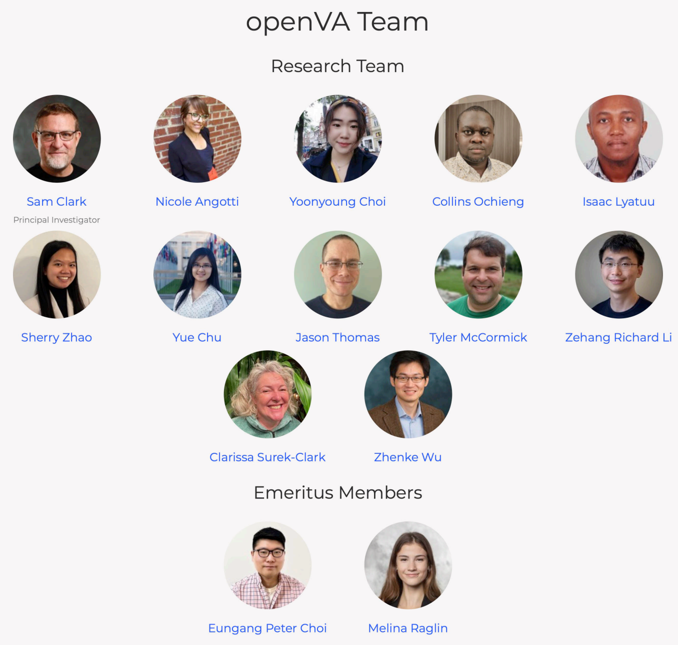
Introduction

Example 1: Age-specific Mortality Models

Example 2: Verbal Autopsy

Closing

openVA Team



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
- info@openva.net
- help@openva.net
- https://samclark.net
- http://openva.net
- https://github.com/verbal-autopsy-software
- These slides: https://samclark.net/biomath2024

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.

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*.

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.

Mikkelsen, L., D. E. Phillips, C. AbouZahr, P. W. Setel, D. De Savigny, R. Lozano, and A. D. Lopez (2015). A global assessment of civil registration and vital statistics systems: monitoring data quality and progress. *The Lancet 386*(10001), 1395–1406.

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.