Verbal Autopsy

Algorithms, Software, & Future Developments

Jason Thomas
Research Scientist &
Software Developer

CHAMPS Analysis Forum March 3rd, 2025

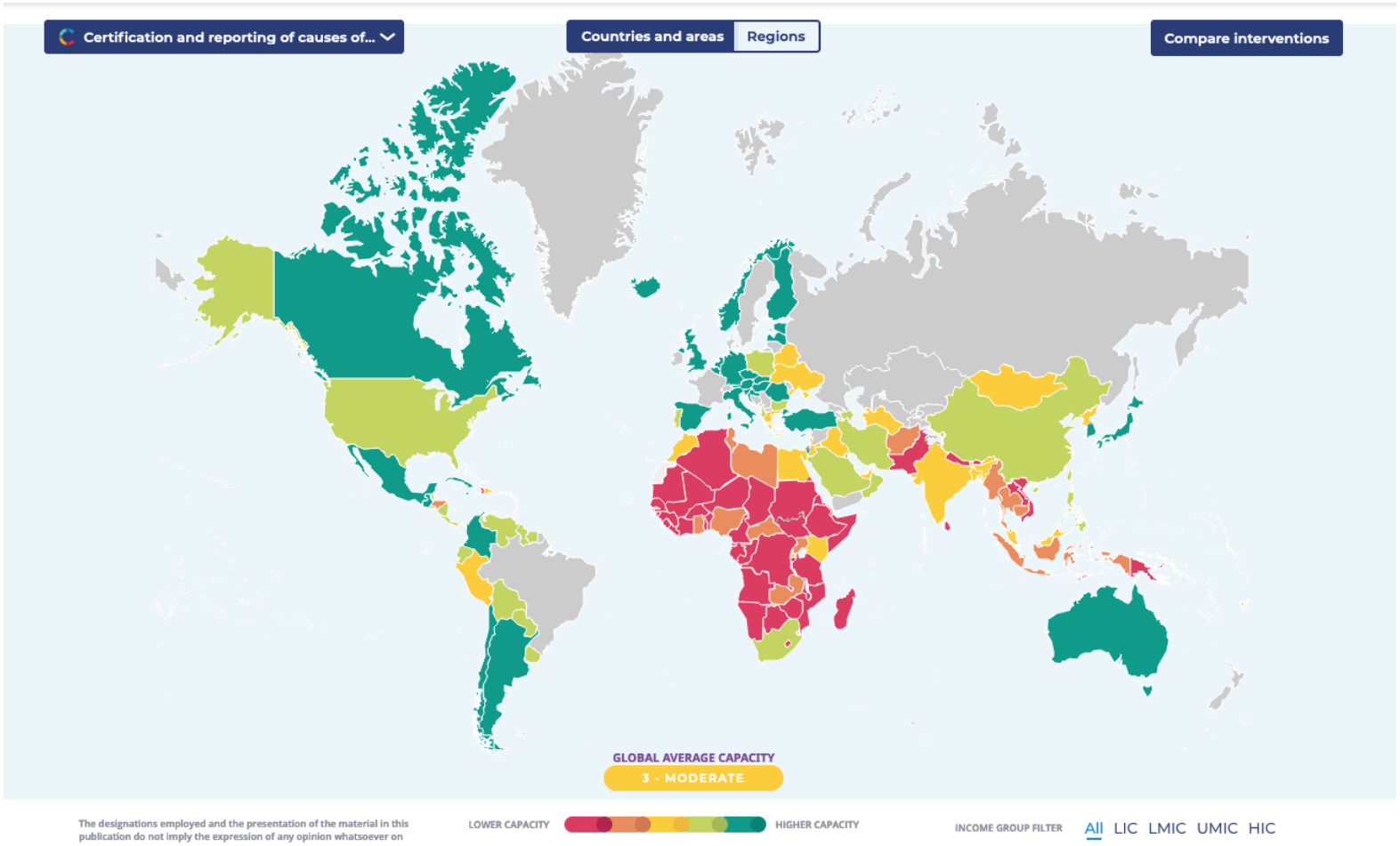
Motivation

Table 1.1 Current status of completeness of death registration and cause of death statistics in Tanzania.

Indicator	Value
Tanzania Population (2021)	57.724 million *
Deaths expected (at a CDR of 6.0/1,000)	346,344*
Deaths with MCCD medically certified cause	51,906 (15%) ** from health facilities
Deaths with MCCD valid for vital statistics	42,142 (12%) **
Deaths registered by CRVS	35,573 (10%) ***
* National Bureau of Statistics Population projection for the mainland in 2021 ** Tanzania MoH from DHIS2 and ANACoD3 *** Tanzania RITA CRVS Authority	

Source: Causes of Deaths From Community Settings in Tanzania. Ministry of Health (2024)

Motivation

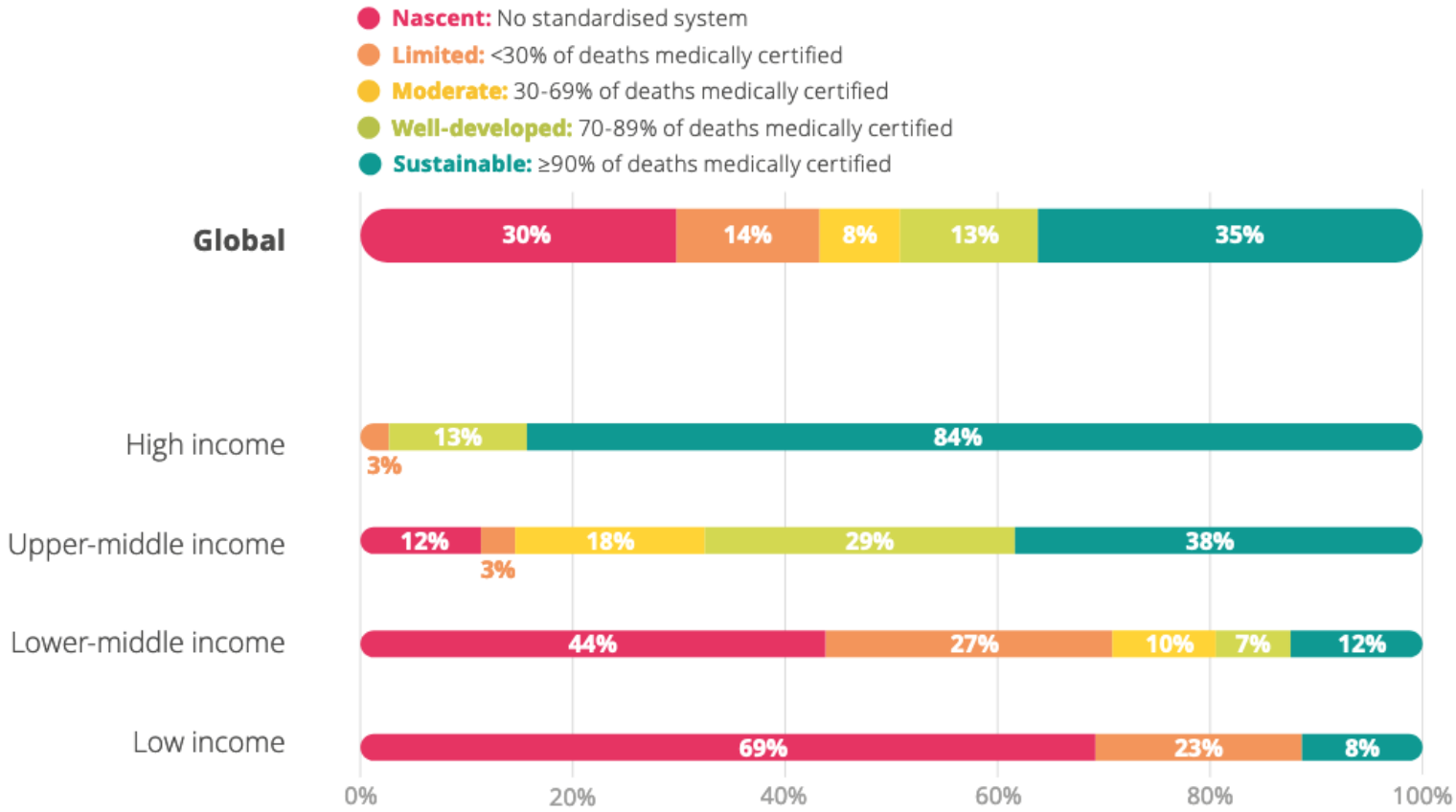


Source: WHO Score Dashboard

Capacity to Register CoD

FIGURE C2.2

PERCENTAGE OF COUNTRIES (N=133), BY CAPACITY TO REGISTER CAUSE OF DEATH, AND COUNTRY INCOME GROUP



Source: "Global report on health data systems and capacity, 2020" WHO

Implications & Solutions

- Knowledge of the leading causes of death at the community/population level informs health polices and facilitates the evaluation of interventions.

Implications & Solutions

- Knowledge of the leading causes of death at the community/population level informs health polices and facilitates the evaluation of interventions.
    - It can take many years for civil registration and vital statistics systems to develop sufficient capacity.

Implications & Solutions

- Knowledge of the leading causes of death at the community/population level informs health polices and facilitates the evaluation of interventions.
    - It can take many years for civil registration and vital statistics systems to develop sufficient capacity.

Verbal Autopsy

interviewing the decedent's next of kin or caregiver to gather data on the relevant symptoms and medical history.

Implications & Solutions

- Knowledge of the leading causes of death at the community/population level informs health polices and facilitates the evaluation of interventions.
    - It can take many years for civil registration and vital statistics systems to develop sufficient capacity.

Verbal Autopsy

interviewing the decedent's next of kin or caregiver to gather data on the relevant symptoms and medical history.

- Physicians can assign causes of death using the VA data.

Implications & Solutions

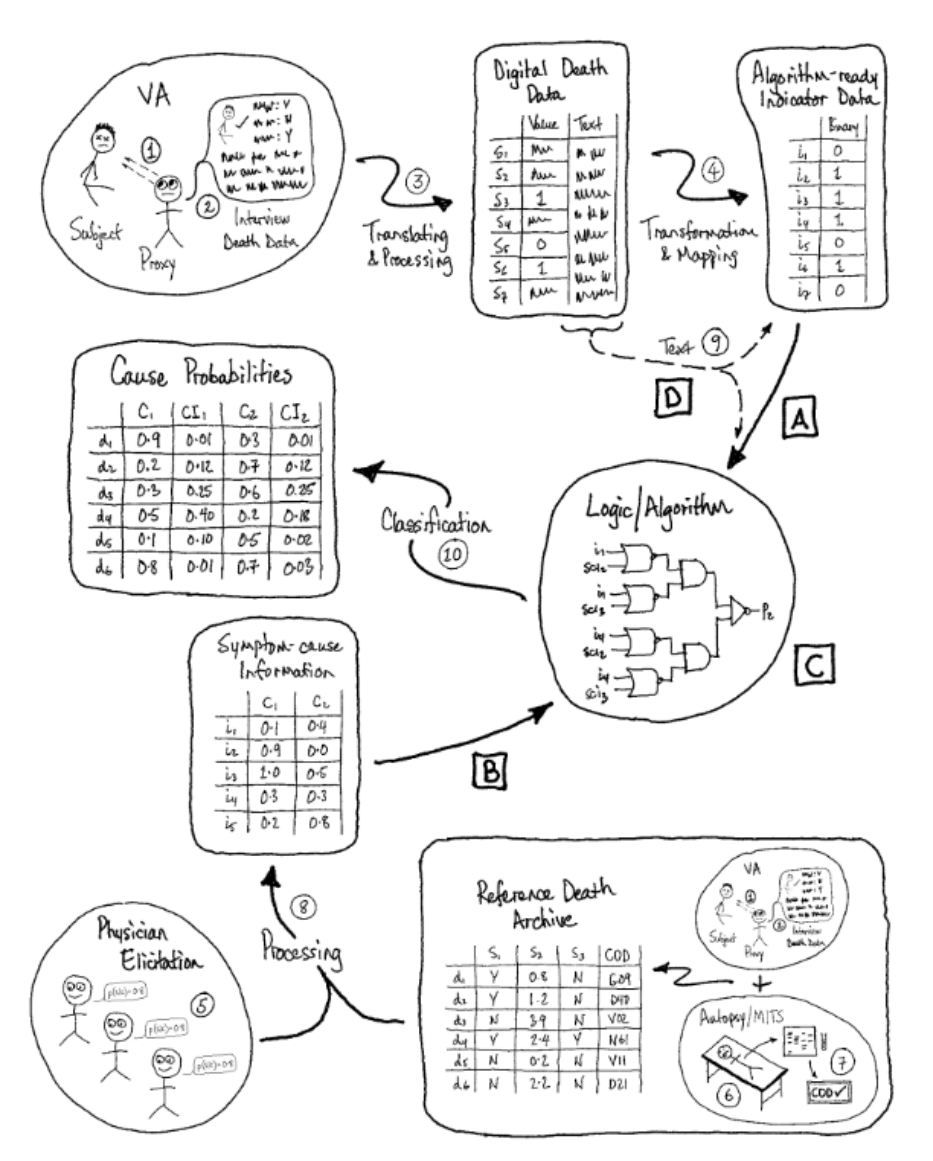
- Knowledge of the leading causes of death at the community/population level informs health polices and facilitates the evaluation of interventions.
    - It can take many years for civil registration and vital statistics systems to develop sufficient capacity.

Verbal Autopsy

interviewing the decedent's next of kin or caregiver to gather data on the relevant symptoms and medical history.

- Physicians can assign causes of death using the VA data.
- Algorithms have also been developed to assign causes, providing an expedient solution for providing CoD information.

VA-verse


VA
Subject
Proxy
Interview
Death Data
Digital Death Data
Translating & Processing
Algorithm-ready Indicator Data
Transformation & Mapping
Text
D
A
Cause Probabilities
Classification
Logic/Algorithm
C
Symptom-cause Information
B
Physician Elicitation
Processing
Reference Death Archive
Autopsy/MITS
COD

openVA Team – openva.net

- The openVA team develops and maintains various tools, algorithms, and software related to Verbal Autopsy.

openVA Team – openva.net

- The openVA team develops and maintains various tools, algorithms, and software related to Verbal Autopsy.
  - team is led by Sam Clark (The Ohio State University)
  - diverse group of statisticians, demographers, sociologists, linguists, computer scientists

openVA Team – openva.net

- The openVA team develops and maintains various tools, algorithms, and software related to Verbal Autopsy.
  - team is led by Sam Clark (The Ohio State University)
  - diverse group of statisticians, demographers, sociologists, linguists, computer scientists
- Primarily produce open-source software packages for R and Python
  - GitHub: https://github.com/verbal-autopsy-software

openVA Team – openva.net

- The openVA team develops and maintains various tools, algorithms, and software related to Verbal Autopsy.
  - team is led by Sam Clark (The Ohio State University)
  - diverse group of statisticians, demographers, sociologists, linguists, computer scientists
- Primarily produce open-source software packages for R and Python
  - GitHub: https://github.com/verbal-autopsy-software
- Research, Training, & User Support
  - info@openva.net & help@openva.net

Workflow – data collection

- Open Data Kit – collect & center
  - Kobo Toolbox is a popular alternative

Workflow – data collection

- Open Data Kit – collect & center
  - Kobo Toolbox is a popular alternative
  - openVA tools assume data are from ODK Central export

Workflow – data collection

- Open Data Kit – collect & center
  - Kobo Toolbox is a popular alternative
  - openVA tools assume data are from ODK Central export
- WHO verbal autopsy instrument
  - openVA Team assisted with the 2022 revision
  - openVA tools process 2012(?), 2016, 2022 instruments

Workflow – data collection

- Open Data Kit – collect & center
  - Kobo Toolbox is a popular alternative
  - openVA tools assume data are from ODK Central export
- WHO verbal autopsy instrument
  - openVA Team assisted with the 2022 revision
  - openVA tools process 2012(?), 2016, 2022 instruments
- Interviewer effects (indeterminate CoD)
  - telaVAs – telephonic verbal autopsies (South Africa)

Workflow – data transformation

Once data have been collected. . .

- data quality & consistency checks

Workflow – data transformation

Once data have been collected. . .

- data quality & consistency checks
- data transformation

pyCrossVA

Python package with functions that convert VA data into the format expected by the algorithms (differences between InterVA & InSilicoVA)

Workflow – data transformation

Once data have been collected. . .

- data quality & consistency checks
- data transformation

pyCrossVA

Python package with functions that convert VA data into the format expected by the algorithms (differences between InterVA & InSilicoVA)

- (also an R package, but it has not been kept up to date)

Workflow – data transformation

Once data have been collected. . .

- data quality & consistency checks
- data transformation

pyCrossVA

Python package with functions that convert VA data into the format expected by the algorithms (differences between InterVA & InSilicoVA)

- (also an R package, but it has not been kept up to date)
- 2022 data with algorithms expecting 2016 format?

Workflow – algorithms

- Data consistency check (vacheck, InterVA5 R package)
- Symptom Cause Information (SCI)
  - **Probbase** – matrix of conditional probabilities: Pr(observing symptom | cause of death)
- Algorithm logic
  - Bayes' Rule
  - Threshold

InterVA

- **Inter**pret **VA** $\rightarrow$ InterVA
- Developed by late Peter Byass and colleagues and refined over many years from roughly 2000 - 2020 (DOI)
- Supports standard WHO VA instruments: 2007, 2012, 2016
- Software available from openVA Team - exactly replicated Byass' software
  - R package Python package
- Probbase: conditional probabilities of observing a VA indicator given a specific cause
  - Elicited directly from physicians, represents physicians' knowledge
  - Does not quantify relationships between groups of causes and causes

InterVA Logic

- Only uses information on VA indicators that are present (for most symptoms); ignores information on VA indicators that are absent (half of Bayes' Rule)
- Epidemiological parameters: level of mortality from HIV and Malaria (set's prior probability)
- For each death, produces individual-level propensity associated with each cause
  - Propensities with values less than 0.4 $\rightarrow$ indeterminate CoD
- Standard software reports top three propensities for each death
- CSMFs calculated by summing up propensities for each cause across all deaths

InSilicoVA

- InSilicoVA $\rightarrow$ in-silicon VA, like in-vivo and in-vitro, i.e. on a computer chip
- Created by Zehang (Richard) Li, Tyler McCormick, and Sam Clark to improve on InterVA by:
  - Utilizing information on both present and absent VA indicators
  - Eliminating undetermined causes
  - Estimating uncertainty/confidence associated with both ICSMs and CSMFs
- Supports standard WHO VAs: 2012, 2016
- Software available from openVA Team
  - R package - continuously maintained
  - Python package

InSilicoVA (cont)

- Produces distributions of cause-specific probabilities for each cause for each death
- Produces distributions of cause-specific mortality fractions for a population of deaths
- Distributions used to identify central probabilities and uncertainty/confidence
- No undetermined causes
  - Easily identified causes have high probability and low uncertainty
  - Hard-to-identify causes have low probability and high uncertainty

Current Developments

2022 WHO VA instrument

- Probbase elicitation
  - started with over 50 physicians from around the world
  - asynchronous rounds & synchronous discussion round (this is where we are now)
- Working on the pyCrossVA mapping for the new Probbase
- New software for the updated algorithms will be ready shortly after
- Original deadline was end of March (going to be pretty close)

Future Developments

There are various ways of improving VA cause assignment using computer coded algorithms

- Improving algorithm performance
  - symptom dependence
  - domain adaptation models
- Use more information from the VA interview
- Growing the body of VA data with reference causes to train algorithms/models
  - build knowledge on epidemiological differences across space & time

Thank you for your time and attention

Any questions???

Misc

Next it is necessary to specify basic epidemiological parameters for two important diseases whose prevalence varies widely from place to place. These are malaria and HIV/AIDS. You can choose either “H” for “high”, “L” for “low” or “V” for “very low” separately for each of these diseases. These settings approximate to “high” being more than 1:100 of all deaths in a population, “low” being around 1:1000 and “very low” being under 1:10000. The default settings are “low”. Examples of appropriate responses might be low malaria, low HIV for many Asian locations; high malaria, high HIV for many East African locations; high malaria, low HIV for some West African locations, etc. The “very low” setting should be used for locations where deaths from malaria or HIV are known to be extremely rare.