DATA NOTE

A de-identified database of 11,979 verbal autopsy open-ended responses [version 1; peer review: 2 approved]

Abraham D. Flaxman, Lisa Harman, Jonathan Joseph, Jonathan Brown, Christopher J.L. Murray

Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, 98121, USA

v1 **First published:** 17 Apr 2018, **2**:18
https://doi.org/10.12688/gatesopenres.12812.1
Latest published: 17 Apr 2018, **2**:18
https://doi.org/10.12688/gatesopenres.12812.1

Abstract

As part of the Gates Grand Challenge 13, the Population Health Metrics Research Consortium (PHMRC) collected data to enable the development and validation of methods that measure cause-specific mortality in populations with incomplete or inadequate cause of death coding.

This work yielded 11,979 verbal autopsy interviews (VAIs). In each, a field interviewer spoke with an individual familiar with the deceased and their final illness, and used a semi-structured questionnaire to collect information about the symptoms of the deceased in their final illness. The VAI collected demographic characteristics, possible risk factors (such as tobacco use), and other potentially contributing characteristics. It also included the open-ended question, *"Could you please summarize, or tell us in your own words, any additional information about the illness and/or death of your loved one?"* (open narrative).

The VAI data were released in a de-identified format in September 2013 through the Global Health Data Exchange, in files that contain verbal autopsies that were collected at six sites in four countries (India, Mexico, Tanzania, and the Philippines).

Due to research interest, we have now created redacted versions of the open narratives from the open-ended question of the questionnaire. We hope that this database will be the source of innovations that increase our knowledge about the causes of ill health and, through this knowledge, produce improvements in health for individuals and populations.

Keywords

Verbal autopsy, cause of death, natural language processing, open data

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 17 Apr 2018	✓ view	✓ view

1. **Michel Garenne**, University of the Witwatersrand, Johannesburg, South Africa

2. **John Hart**, London School of Hygiene and Tropical Medicine, London, UK
The University of Melbourne, Melbourne, Australia

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Abraham D. Flaxman (abie@uw.edu)

Author roles: Flaxman AD: Conceptualization, Data Curation, Investigation, Methodology, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Harman L**: Data Curation, Investigation, Writing – Review & Editing; **Joseph J**: Data Curation, Investigation, Writing – Review & Editing; **Brown J**: Data Curation, Project Administration, Supervision, Writing – Review & Editing; **Murray CJL**: Conceptualization, Funding Acquisition, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by a grant from the Bill and Melinda Gates Foundation through the Grand Challenges in Global Health initiative [OPP37883]. The funders had no role in study design, data collection and analysis, interpretation of data, decision to publish, or preparation of the manuscript. The corresponding author had full access to all data analyzed and had final responsibility for the decision to submit this original research paper for publication.

How to cite this article: Flaxman AD, Harman L, Joseph J *et al.* **A de-identified database of 11,979 verbal autopsy open-ended responses [version 1; peer review: 2 approved]** Gates Open Research 2018, **2**:18 https://doi.org/10.12688/gatesopenres.12812.1

First published: 17 Apr 2018, **2**:18 https://doi.org/10.12688/gatesopenres.12812.1

Introduction

Population health information that is both accurate and comprehensive can aid program implementation, monitoring, and evaluation, resource allocation and planning. However, there are currently large gaps in the technologies and measurement methods that are available to generate this information, and this makes it difficult to address health inequities through effective policy[1].

The Population Health Metrics Research Consortium (PHMRC) conducted data collection to enable the development and validation of methods that measure cause-specific mortality in populations with incomplete or inadequate cause of death coding. This work produced around 12,000 verbal autopsy interviews (VAIs), in which a relative or someone familiar with the final illness of the deceased, provides information about the signs symptoms of the final illness, as well as demographic characteristics, and information on risk factor exposures (such as tobacco use), and other potentially relevant characteristics[2].

The VAI data were released in a de-identified format in September 2013, through the Global Health Data Exchange, in files that contain verbal autopsies from six sites in four countries (India, Mexico, Tanzania, and the Philippines) using a standardized VA questionnaire developed by the PHMRC. The data is organized into three parts corresponding to the questionnaire modules for each age group: neonate, child, and adult. Each VAI in the database is matched with a "gold standard" diagnoses of underlying causes of death, typically identified from medical records, and using stringent diagnostic criteria (such as laboratory, pathology, or medical imaging findings.)[3]

One portion of a VAI is the "open narrative," where the respondent has the opportunity to tell, in their own words, what happened during the illness that led to the death being investigated. This was collected as a final question in the PHMRC survey, after the structured interview, when the respondent was asked, *"Could you please summarize, or tell us in your own words, any additional information about the illness and/or death of your loved one?"* The full response to this question was transcribed and translated into English, and the 2013 data release included counts of stemmed keywords as variables in the final dataset, to allow researchers access to this rich source of unstructured data, while also removing any potentially personally identifiable information (PII) in that portion of the interview.

Due to research interest, we have now created redacted versions of 11,979 open narratives to allow researchers the opportunity to learn even more about how deaths are described. We hope that this database will be the source of innovations that increase our knowledge about the causes of ill health and through knowledge produce improvements in health for individuals and populations.

Methods

The process of collecting the VAIs has been described in detail previously[1]. In this article, we provide a detailed account of the protocol used to redact personal information from the open-ended question, and therefore allow the release of the full text of the open narrative collected in the VAIs.

Study participants provided their consent to participate with the knowledge that "reports of the data … will not identify any individual person." We chose also to redact the names of specific health facilities to avoid the risk of identifying individual health service providers indirectly, through their association with individual facilities. To retain the most information possible for future research, we replaced PII with "tags" that denote what sort of information has been redacted.

An example makes this clear: a typical text was redacted to read, "vaginal bleeding and delay to receive care at [HOSPITAL] was the main cause of death. he said that his wife arrive at the hospital at 8pm and didn't receive any care until 8am." Instead of including the name of the specific hospital, we redacted it to [HOSPITAL]. The tags used to replace PII are [HOSPITAL], [DOCTOR], [PATIENT], [PLACE], [PERSON], and [YEAR].

We initially planned to redact dates entirely but chose to redact only the year, to make it easier for future researchers to measure the time between events. To allow for different years, we used the tag [YEAR + n]. An example is "last november of [YEAR]-the deceased got stroke left side of his body. was hospitalized due to high blood pressure last year. january this year was his last hospitalization that leads to death. jan. 26, [YEAR+1]. experienced fast breathing, that's why he was brought to the hospital (provincial hospital). with oxygen and ngt; got fever and cough; in coma. jan. 31, [YEAR+1]. was tried to revive around 11:00 pm to 12 midnight, but was not able to revive him. around 3:00 am (at dawn), he died."

When a response referred to multiple different specific hospitals, we redacted the hospitals to [HOSPITAL] and [HOSPITAL2]. Subsequent distinct hospitals in same passage were redacted to [HOSPITAL3], [HOSPITAL4], etc.

We included all VAIs for which there was an open-response string available to redact, even when the response was devoid of information.

We implemented the redaction process in a spreadsheet using Excel 2010, redacted manually by a single data analyst (LH), who read each open-response and replaced each piece of PII with the appropriate tag.

Redaction rules, with some examples and counterexamples

1. Specific patient becomes [PATIENT]

   Example: John Smith was taken to … -> [PATIENT] was taken to

   Counterexamples (no redaction for the following): She was taken to -> She was taken to (no change)

   My uncle was taken to -> My uncle was taken to (no change)

   The patient was taken to -> The patient was taken to (no change)

2. Specific health facility becomes [HOSPITAL],
3. Specific doctor becomes [DOCTOR]
4. Specific place that is not a health facility becomes [PLACE]
5. Specific person that is not doctor or patient becomes [PERSON]

Iterative development of redaction rules

We originally planned to redact date (including day, month, and year) to [DATE], but to maintain time sequence, we changed this to not redact entire date where month and/or day show time progression. Only redact [YEAR] to keep the reference to time elapsed. See example below where specific dates show progression of time.

Examples of [YEAR] redactions

jan. 12, [YEAR]. she was bumped by a motorcycle which seened like it had no lights. the deceased had a little drink at that time and her sense of hearing was poor. she was going to cross the street when that happened. she was brought to the hospital but she was unconscious. her breathing was controlled by a pump. the accident happened at around 6 pm jan. 13, [YEAR]. at around 6 am we found out she's dead because the cardiac monitor showed a straight line.

If progression of time spans over years, [YEAR+n] is used. See example below, where passage refers to following year:

last november of [YEAR]-the deceased got stroke left side of his body. was hospitalized due to high blood pressure last year. january this year was his last hospitalization that leads to death. jan. 26, [YEAR+1]. experienced fast breathing, that's why he was brought to the hospital (provincial hospital). with oxygen and ngt; got fever and cough; in coma. jan. 31, [YEAR+1]. was tried to revive around 11:00 pm to 12 midnight, but was not able to revive him. around 3:00 am (at dawn), he died.

Where a passage refers to two different hospitals, hospitals are redacted to [HOSPITAL] and [HOSPITAL2]. Subsequent hospitals in same passage would be [HOSPITAL3, [HOSPITAL4], etc:

may 16, [YEAR]. he got accident. was brought immediately to [HOSPITAL] then referred directly to [HOSPITAL2], there his wound was stitched. his head was the affected part. was referred to [HOSPITAL3]. was ct scanned in [HOSPITAL4], there was a break on his forehead. was operated after 2 days. after operation he got fever. the deceased also had cough. as per respondent, it was not just the accident alone who led the deceased to death. there was also a complication of his kidney disease. long before (respondent was not able to remember the exact date), the deceased experienced inability to walk but it was not consulted to the doctor for the deceased doesn't want to. they only went to a traditional healer for treatment. the deceased can't walk for about 7 months but then later on he was able to walk again. after he was also hospitalized at [HOSPITAL5], it was known that he have kidney disease.

Additional clarifications

Midwife names were redacted to [DOCTOR].

Dataset validation

We reviewed progress weekly and discussed emerging challenges as they arose. For example, we determined that the original plan of redacting dates entirely to [DATE] seemed to be obscuring valuable information about the time between symptoms. One week later, we determined that our first attempt at a remedy, to include [DATE+days] was to labor intensive, and would prevent redaction from completing within our budget. Our next remedy worked, and that is how we developed the [YEAR+n] approach described above. When redaction was completed, we reviewed a simple random sample of redacted texts and confirmed that all were devoid of PII.

Ethics approval

This study was approved by the Human Subjects Division of the University of Washington (application number 34413). Ethical approval sought for the VAIs is stated in 1. All data were collected with informed verbal consent from participants before beginning the interview.

Data availability

Data underlying the study are available on OSF: http://doi.org/10.17605/OSF.IO/XUK5Q[4]

Competing interests

No competing interests were disclosed.

Grant information

This work was funded by a grant from the Bill and Melinda Gates Foundation through the Grand Challenges in Global Health initiative [OPP37883].

The funders had no role in study design, data collection and analysis, interpretation of data, decision to publish, or preparation of the manuscript. The corresponding author had full access to all data analyzed and had final responsibility for the decision to submit this original research paper for publication.

Acknowledgements

The authors would like to thank Scott Lee for the data request that first demonstrated demand for making the data public.

References

1. Mikkelsen L, Phillips DE, AbouZahr C, *et al.*: **A global assessment of civil registration and vital statistics systems: monitoring data quality and progress.** *Lancet.* 2015; **386**(10001): 1395–406.
PubMed Abstract

2. Murray CJ, Lopez AD, Black R, *et al.*: **Population Health Metrics Research Consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets.** *Popul Health Metr.* BioMed Central; 2011; **9**: 27.
PubMed Abstract | Publisher Full Text | Free Full Text

3. **Population health metrics research consortium gold standard verbal autopsy data 2005–2011**. Population Health Metrics Research Consortium (PHMRC); 2013.
Reference Source

4. Flaxman AD: "**A de-Identified Database of 11,979 Verbal Autopsy Open-Ended Responses.**" *Open Science Framework.* 2018.
Data Source

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 14 June 2018

https://doi.org/10.21956/gatesopenres.13879.r26471

✔ **John Hart**

[1] Department of Clinical Research, London School of Hygiene and Tropical Medicine, London, UK
[2] Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Vic, Australia

The authors present the release of narrative free text from verbal autopsy (VA) interview data collected as part of the PHMRC gold standard VA validation study. The methods describe the rules and processes for redacting personally identifiable information.

The release of the data is welcome as including the open narrative in automated VA analysis programs can significantly alter output diagnoses. Better understanding of the pros and cons of using free text, and the appropriate weighting for different sources of free text, is likely to be of further interest to researchers as their methods and technology develop. To illustrate, simple text mining for key words such as "malaria" could correctly provide positive evidence towards a diagnosis if the free text read "She suffered from malaria..." but incorrectly if the free text read "Her malaria tests at the health centre were negative".

Research may use the narrative free text alone but its use in VA analysis algorithms is likely to complement answers to the structured interview component of VA. This dataset includes the gold standard diagnosis and narrative free text but no link to the structured answers. The authors may wish to comment on this.

The methods are well described, with appropriate examples, and the released dataset clear to understand whilst minimising risk of individual identification.

Typos
Page 3: "Each VAI in the database is matched with a “gold standard” diagnoses of underlying causes of death" rather "Each VAI in the database is matched with a “gold standard” diagnosis of underlying cause of death".

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 07 June 2018

https://doi.org/10.21956/gatesopenres.13879.r26486

Michel Garenne
MRC/Wits Rural Public Health and Health Transitions Research Unit, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

This note presents a new database of Verbal Autopsy (VA) narratives gathered by the Population Health Metrics project. The database is placed in open access and is easily accessible; it ensures the confidentiality of the persons who answered the questionnaire.

This database has a rich potential for further research. Firstly, it can be used for analyzing what caught people's eye: signs, symptoms and circumstances that people tend report in specific circumstances; for analyzing what was ignored or remained unreported although most likely present according to the medical diagnosis; and for determining who knows about the precise cause of death learned from the medical authorities. This information could be used for better understanding VA narratives. Secondly, it could be used for further refining the VA diagnoses, although this was already done in parts by detecting key words in earlier studies. Thirdly, the database can be used for a full scale textual analysis, including style of reporting, keywords, underlying emotions, selectivity, etc. Lastly, relating the full scale textual analysis with the medical diagnosis and with the answers to the VA structured questionnaire could be most useful for further research.

This database is therefore most welcome, and likely to become a source of numerous research exercises.

Typos:

1. Page 4: "*she was bumped by a motorcycle which seened like it had no lights".* Rather "seemed"

2. In the data file, the codebook for Study Site is duplicated.

3. There are many typos in the narratives (example: diabtes for diabetes, etc.). Could this be cleaned? This would help for searching key words and performing textual analysis.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Comments on this article

Version 1

Reader Comment 24 Apr 2018

Peter Byass, Umeå University, Sweden

The publication of these free-text narrative portions of the Population Health Metrics Research Consortium verbal autopsy dataset is very welcome, complementing the earlier publication of responses to the hundreds of individual responses to closed questions for each case.

However, much of the real scientific potential in publishing these free-text narratives would lie in enabling analyses on a case-by-case basis of the free text against the closed-question responses. However, since there does not appear to be any common anonymous case identifier linking the previously published closed-question responses to the these free-text narratives, most of the scientific potential in this publication is lost.

Competing Interests: no competing interests