Determining Causes of Death from Verbal Autopsy with ChatGPT: A Case Study of 6942 Deaths in Sierra Leone from 2019-2022

Richard Wen[1*], Rajeev Kamadod[1,5], Cheryl Chin[1], Asha Behdinan[1,3], Leslie Newcombe[1], Anteneh Tesfaye Assalif[1], Areeba Zubair[1,3], Thomas Kai Sze Ng[1], Patrick Brown[1,4], Prabhat Jha[1,2]

[1*]Centre for Global Health Research, St. Michael's Hospital, Unity Health Toronto, 30 Bond St, Toronto, M5B 1W8, Ontario, Canada.
[2]Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, M5T 3M7, Ontario, Canada.
[3]Department of Surgery, University of Toronto, 149 College Street, Toronto, M5T 1P5, Ontario, Canada.
[4]Department of Statistical Sciences, University of Toronto, 700 University Avenue, Toronto, M5G 1Z5, Ontario, Canada.
[5]Kentropy Technologies Pvt. Ltd., 2nd Main Rd, Bengaluru, 560 034, Bangalore, India.

*Corresponding author(s). E-mail(s): richard.wen@unityhealth.to; Contributing authors: rajeevk@kentropy.com; cheryl.chin@unityhealth.to; asha.behdinan@mail.utoronto.ca; leslie.newcombe@unityhealth.to; antenehta@gmail.com; areeba.zubair@mail.utoronto.ca; KaiSze.Ng@unityhealth.to; ; prabhat.jha@utoronto.ca;

Abstract

The abstract serves both as a general introduction to the topic and as a brief, non-technical summary of the main results and their implications. Authors are advised to check the author instructions for the journal they are submitting to for word limits and if structural elements like subheadings, citations, or equations are permitted.

Keywords: keyword1, Keyword2, Keyword3, Keyword4

1 Background

In 2019, 41 million people die prematurely from noncommunicable diseases every year, accounting for 74% of all deaths globally [1]. Most of these deaths are preventable, but require adequate resource allocation, guided by evidence, to implement effective interventions and policies that target populations at risk [2]. Thus, reliable counts and diagnoses of deaths provide decision makers with evidence to save lives and reduce premature deaths worldwide [3–6]. However, most low-income countries do not have data on deaths or have registered less than half of the deaths in their country, with an even fewer 8% of these registered deaths having a Cause of Death (COD) recorded [7]. To fill this gap in death registrations, an alternative method known as Verbal Autopsy (VA) is used to collect data on deaths and determine their likely causes at scale [8–10], outside of traditional healthcare facilities where over half of deaths occur at home [11].

VA involves two major components: survey and COD assignment [12, 13]. In the survey component, trained lay surveyors interview those familiar with the deceased (e.g. living spouse, children, family, friends) to gather information using standardized questionnaires and open narratives. In the COD assignment component, physicians evaluate information available from the questionnaires and open narratives to assign probable CODs. Although VA surveys have been an effective alternative to collect mortality data at scale (e.g. less than $3 USD per house in India [14, 15]), COD assignment has been criticized to be expensive and difficult to reproduce due to reliance on physician assignment [16, 17]. As an alternative to physician assignment, computer models, such as InterVA [18] and InSilicoVA [16], have recently been studied to automatically assign CODs with performances close to physicians at the population level, but poor performances at the individual level [19–22]. These computer models often utilized data from the structured questionnaire, but often omit the free-text open narrative, which misses latent information, such as chronology or health-seeking behaviors, that may potentially help models perform better than using the questionnaire alone [23–25].

Recently, Large Language Models (LLM), leveraging massive datasets and deep learning approaches, have made advances in performing a variety of Natural Language Processing (NLP) tasks using free-text, such as question answering, code generation, and even medical diagnosis [26–29]. On November 30, 2022, a widely-available LLM called ChatGPT was released by OpenAI with capabilities of answering natural language text inquiries using training data up to September 2021. ChatGPT-3 was based on several Generative Pre-trained Transformer (GPT) models between 2018 to 2020, namely GPT-1 to GPT-3, which had notable differences in training data sizes of 5 gigabytes to 45 terabytes from web sources that resulted in 117 million to 175 billion parameter models [30]. On March 14, 2023, ChatGPT-4 was released with human-level performance on various professional and academic exams and benchmarks that outperformed ChatGPT-3 [31]. Given the limited usage of free-text open narratives in computer models for determining CODs, and recent advances in LLMs that leverage natural language text prompts, we conduct a case study with Sierra Leone deaths from VA in 2019 to 2022 to compare the performances of four models, ChatGPT-3.5, ChatGPT-4, InterVA-5, and InSilicoVA, for determining CODs.

2 Methods

The performances of four models, ChatGPT-3.5, ChatGPT-4, InterVA-5, and InSilicoVA, for determining CODs were evaluated using 6939 physician agreed records in 2019 to 2022 from the Health Sierra Leone (HEAL-SL) study [32, 33]. Physician agreed records are dual-coded records where two physicians had similar COD assignments. COD outputs from the four models were compared to CODs assigned by physicians.

2.1 Physician Agreed Records

Initially, 11,920 records were collected from dual-coded Electronic Verbal Autopsy (EVA), where each record was randomly coded by two different physicians that assigned CODs as International Classification of Diseases Revision 10 (ICD-10) codes [34]. Physicians were able to assign CODs for 11,820 of the 11,920 records, where 100 of these records could not be assigned a COD due to missing or inadequate information (e.g. low quality narrative, data loss). To determine if two codes were in physician agreement, codes from two physicians per record were compared in the EVA system to determine if they agreed using Central Medical Evaluation Agreement 10 (CMEA-10) codes, which groups a range of similar ICD-10 codes together [35] (see Additional File 1). When codes were not in agreement, a record enters the reconciliation phase, where the two physicians were provided reasoning and codes from each other to: (1) keep their initial code (2) assign the other physician's code or (3) assign a new code. If codes are not in agreement after the reconciliation phase, a record enters the adjudication phase, where a third senior physician evaluates both physicians' reasoning and codes, and assigns a final code. The 11,820 physician coded records were further filtered for records where both physicians agreed on the assigned codes (records that were not reconciled or adjudicated), resulting in 6942 physician agreed records. Since computer models were compared to physicians, there was more certainty that COD assignments agreed by both physicians were representative of physician assignment than when they disagreed [17, 36, 37].

2.2 Cause of Death Prediction Models

Four computer models were used to predict COD for each of the 6942 physician agreed records: ChatGPT-3.5 [38], ChatGPT-4 [31], InterVA-5 [18], and InSilicoVA [16]. Each model required pre-processing of the 6942 physician agreed records into input data, and standardization of output COD codes from models for performance evaluation as not all models produced comparable codes across outputs.

2.2.1 Input Data

For ChatGPT-3.5 and ChatGPT-4, prompts were generated for each record as input data with a template involving `<narrative>` from the free-text open narratives, and associated `<age>`, (age) `<unit>`, `<sex>` from the standardized questionnaire:

```
Determine the underlying cause of death and provide a ICD−10 code
for the following narrative based on a verbal autopsy of a death in
Sierra Leone, Age <age> <unit>, <sex>: <narrative>
```

For InterVA-5 and InSilicoVA, the standardized questionnaire data from the HEAL-SL EVA were first converted into 2016 World Health Organization (WHO) VA questionnaire revision 1.5.1 Open Data Kit (ODK) format [39]. The WHO revision 1.5.1 ODK data were further converted into OpenVA format [40] using the `pyCrossVA` Python package [41], and used as input for InterVA-5 and InSilicoVA. The input data were 6939 generated prompts for ChatGPT-3.5 and ChatGPT-4, and X OpenVA formatted records for InterVA-5 and InSilicoVA.

2.2.2 Output Data

ChatGPT-3.5, ChatGPT-4, InterVA-5, and InSilicoVA were able to output COD codes for X, X, X, X of the X records respectively. Thus, only X records with an output code for all four models were kept for evaluation. ChatGPT-3.5 and ChatGPT-4 output ICD-10 codes, but InterVA-5 and InSilicoVA output WHO VA 2016 codes, which were converted into ICD-10 codes. All model output ICD-10 codes were further converted to Centre for Global Health Research 10 (CGHR-10) codes that generalized ICD-10 codes into 19, 10, and 7 categories for the adult (12 years or older), child (28 days to 11 years), and neonatal (under 28 days) age groups (see Additional file 2). X records containing CGHR-10 codes for each of the four models were used for performance evaluation.

2.3 Performance Evaluation

x.

3 Results

x.

4 Discussion

Discussions should be brief and focused. In some disciplines use of Discussion or 'Conclusion' is interchangeable. It is not mandatory to use both. Some journals prefer a section 'Results and Discussion' followed by a section 'Conclusion'. Please refer to Journal-level guidance for any specific requirements.

5 Conclusion

Conclusions may be used to restate your hypothesis or research question, restate your major findings, explain the relevance and the added value of your work, highlight any limitations of your study, describe future directions for research and recommendations.

In some disciplines use of Discussion or 'Conclusion' is interchangeable. It is not mandatory to use both. Please refer to Journal-level guidance for any specific requirements.

Supplementary information. Additional files were used to supplement this paper:

- Additional file 1: Central Medical Evaluation Agreement 10 (CMEA-10) codes. ICD-10 code ranges considered in physician agreement. (.csv)
- Additional file 2: Centre for Global Health Research 10 (CGHR-10) codes. Codes grouping ICD-10 code ranges into generalized categories. (.csv)

Acknowledgments.

Declarations

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article (and its additional files).

References

[1] World Health Organization.: Non Communicable Diseases: Key Facts. Available from: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

[2] Benziger CP, Roth GA, Moran AE. The Global Burden of Disease Study and the Preventable Burden of NCD. Global Heart. 2016 Dec;11(4):393–397. https://doi.org/10.1016/j.gheart.2016.10.024.

[3] Lawn JE, Kerber K, Enweronu-Laryea C, Cousens S. 3.6 Million Neonatal Deaths—What Is Progressing and What Is Not? Seminars in Perinatology. 2010 Dec;34(6):371–386. https://doi.org/10.1053/j.semperi.2010.09.011.

[4] Lassi ZS, Bhutta ZA. Community-based Intervention Packages for Reducing Maternal and Neonatal Morbidity and Mortality and Improving Neonatal Outcomes. Cochrane Database of Systematic Reviews. 2015;(3). https://doi.org/10.1002/14651858.CD007754.pub3.

[5] Liu NH, Daumit GL, Dua T, Aquila R, Charlson F, Cuijpers P, et al. Excess Mortality in Persons with Severe Mental Disorders: A Multilevel Intervention Framework and Priorities for Clinical Practice, Policy and Research Agendas. World Psychiatry. 2017;16(1):30–40. https://doi.org/10.1002/wps.20384.

[6] Ewig S, Torres A. Community-Acquired Pneumonia as an Emergency: Time for an Aggressive Intervention to Lower Mortality. European Respiratory Journal. 2011 Aug;38(2):253–260. https://doi.org/10.1183/09031936.00199810.

[7] World Health Organization. SCORE for Health Data Technical Package: Global Report on Health Data Systems and Capacity, 2020; 2021. Available from: https://www.who.int/publications/i/item/9789240018709.

[8] de Savigny D, Riley I, Chandramohan D, Odhiambo F, Nichols E, Notzon S, et al. Integrating Community-Based Verbal Autopsy into Civil Registration and Vital Statistics (CRVS): System-Level Considerations. Global Health Action. 2017 Jan;10(1):1272882. https://doi.org/10.1080/16549716.2017.1272882.

[9] Thomas LM, D'Ambruoso L, Balabanova D. Verbal Autopsy in Health Policy and Systems: A Literature Review. BMJ Global Health. 2018 May;3(2):e000639. https://doi.org/10.1136/bmjgh-2017-000639.

[10] Rampatige R, Mikkelsen L, Hernandez B, Riley I, Lopez AD. Systematic Review of Statistics on Causes of Deaths in Hospitals: Strengthening the Evidence for Policy-Makers. Bulletin of the World Health Organization. 2014 Sep;92:807–816. https://doi.org/10.2471/BLT.14.137935.

[11] Adair T. Who Dies Where? Estimating the Percentage of Deaths That Occur at Home. BMJ Global Health. 2021 Sep;6(9):e006766. https://doi.org/10.1136/bmjgh-2021-006766.

[12] World Health Organization. Verbal Autopsy Standards: 2022 WHO Verbal Autopsy Instrument; 2023. Available from: https://www.who.int/publications/m/item/training-curriculum-for-the-training-of-verbal-autopsy-master-trainers-and-supervisors.

[13] Chandramohan D, Fottrell E, Leitao J, Nichols E, Clark SJ, Alsokhn C, et al. Estimating Causes of Death Where There Is No Medical Certification: Evolution and State of the Art of Verbal Autopsy. Global Health Action. 2021 Oct;14(sup1):1982486. https://doi.org/10.1080/16549716.2021.1982486.

[14] Gomes M, Begum R, Sati P, Dikshit R, Gupta PC, Kumar R, et al. Nationwide Mortality Studies To Quantify Causes Of Death: Relevant Lessons From India's Million Death Study. Health Affairs. 2017 Nov;36(11):1887–1895. https://doi.org/10.1377/hlthaff.2017.0635.

[15] Jha P, Gajalakshmi V, Gupta PC, Kumar R, Mony P, Dhingra N, et al. Prospective Study of One Million Deaths in India: Rationale, Design, and Validation Results. PLOS Medicine. 2005 Dec;3(2):e18. https://doi.org/10.1371/journal.pmed.0030018.

[16] McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark SJ. Probabilistic Cause-of-Death Assignment Using Verbal Autopsies. Journal of the American Statistical Association. 2016 Jul;111(515):1036–1049. https://doi.org/10.1080/01621459.2016.1152191.

[17] Morris SK, Bassani DG, Kumar R, Awasthi S, Paul VK, Jha P. Factors Associated with Physician Agreement on Verbal Autopsy of over 27000 Childhood Deaths in India. PloS one. 2010;5(3):e9583.

[18] Byass P, Hussain-Alkhateeb L, D'Ambruoso L, Clark S, Davies J, Fottrell E, et al. An Integrated Approach to Processing WHO-2016 Verbal Autopsy Data: The InterVA-5 Model. BMC Medicine. 2019 May;17(1):102. https://doi.org/10.1186/s12916-019-1333-6.

[19] Jha P, Kumar D, Dikshit R, Budukh A, Begum R, Sati P, et al. Automated versus Physician Assignment of Cause of Death for Verbal Autopsies: Randomized Trial of 9374 Deaths in 117 Villages in India. BMC Medicine. 2019 Jun;17(1):116. https://doi.org/10.1186/s12916-019-1353-2.

[20] Leitao J, Desai N, Aleksandrowicz L, Byass P, Miasnikof P, Tollman S, et al. Comparison of Physician-Certified Verbal Autopsy with Computer-Coded Verbal Autopsy for Cause of Death Assignment in Hospitalized Patients in Low- and Middle-Income Countries: Systematic Review. BMC Medicine. 2014 Feb;12(1):22. https://doi.org/10.1186/1741-7015-12-22.

[21] Desai N, Aleksandrowicz L, Miasnikof P, Lu Y, Leitao J, Byass P, et al. Performance of Four Computer-Coded Verbal Autopsy Methods for Cause of Death Assignment Compared with Physician Coding on 24,000 Deaths in Low- and Middle-Income Countries. BMC Medicine. 2014 Feb;12(1):20. https://doi.org/10.1186/1741-7015-12-20.

[22] Tunga M, Lungo J, Chambua J, Kateule R. Verbal Autopsy Models in Determining Causes of Death. Tropical Medicine & International Health. 2021;26(12):1560–1567. https://doi.org/10.1111/tmi.13678.

[23] Jeblee S, Gomes M, Jha P, Rudzicz F, Hirst G. Automatically Determining Cause of Death from Verbal Autopsy Narratives. BMC Medical Informatics and Decision Making. 2019 Jul;19(1):127. https://doi.org/10.1186/s12911-019-0841-9.

[24] Blanco A, Pérez A, Casillas A, Cobos D. Extracting Cause of Death From Verbal Autopsy With Deep Learning Interpretable Methods. IEEE Journal of Biomedical and Health Informatics. 2021 Apr;25(4):1315–1325. https://doi.org/10.1109/JBHI.2020.3005769.

[25] King C, Zamawe C, Banda M, Bar-Zeev N, Beard J, Bird J, et al. The Quality and Diagnostic Value of Open Narratives in Verbal Autopsy: A Mixed-Methods Analysis of Partnered Interviews from Malawi. BMC Medical Research Methodology.

2016 Feb;16(1):13. https://doi.org/10.1186/s12874-016-0115-5.

[26] Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al.: A Survey on Evaluation of Large Language Models. arXiv. Available from: http://arxiv.org/abs/2307.03109.

[27] Lund BD, Wang T. Chatting about ChatGPT: How May AI and GPT Impact Academia and Libraries? Library Hi Tech News. 2023 Jan;40(3):26–29. https://doi.org/10.1108/LHTN-01-2023-0009.

[28] Svyatkovskiy A, Deng SK, Fu S, Sundaresan N. IntelliCode Compose: Code Generation Using Transformer. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ESEC/FSE 2020. New York, NY, USA: Association for Computing Machinery; 2020. p. 1433–1443. Available from: https://doi.org/10.1145/3368089.3417058.

[29] Haupt CE, Marks M. AI-Generated Medical Advice—GPT and Beyond. JAMA. 2023 Apr;329(16):1349–1350. https://doi.org/10.1001/jama.2023.5321.

[30] Wu T, He S, Liu J, Sun S, Liu K, Han QL, et al. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. IEEE/CAA Journal of Automatica Sinica. 2023;10(5):1122–1136. https://doi.org/10.1109/JAS.2023.123618.

[31] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al.: GPT-4 Technical Report. arXiv. Available from: http://arxiv.org/abs/2303.08774.

[32] Njala University.: Healthy Sierra Leone. Available from: https://healsl.org/.

[33] Carshon-Marsh R, Aimone A, Ansumana R, Swaray IB, Assalif A, Musa A, et al. Child, Maternal, and Adult Mortality in Sierra Leone: Nationally Representative Mortality Survey 2018–20. The Lancet Global Health. 2022 Jan;10(1):e114–e123. https://doi.org/10.1016/S2214-109X(21)00459-9.

[34] World Health Organization. ICD-10: International Statistical Classification of Diseases and Related Health Problems (10th Revision); 2011.

[35] Aleksandrowicz L, Malhotra V, Dikshit R, Gupta PC, Kumar R, Sheth J, et al. Performance Criteria for Verbal Autopsy-Based Systems to Estimate National Causes of Death: Development and Application to the Indian Million Death Study. BMC Medicine. 2014 Feb;12(1):21. https://doi.org/10.1186/1741-7015-12-21.

[36] Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. JAMA Network Open. 2019 Mar;2(3):e190096. https://doi.org/10.1001/

jamanetworkopen.2019.0096.

[37] Hsiao M, Morris SK, Bassani DG, Montgomery AL, Thakur JS, Jha P. Factors Associated with Physician Agreement on Verbal Autopsy of over 11500 Injury Deaths in India. PLOS ONE. 2012 Jan;7(1):e30336. https://doi.org/10.1371/journal.pone.0030336.

[38] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al.: Language Models Are Few-Shot Learners. arXiv. Available from: http://arxiv.org/abs/2005.14165.

[39] Nafundi.: ODK - Collect Data Anywhere. Available from: https://getodk.org.

[40] Li ZR, Thomas J, Choi E, McCormick TH, Clark SJ. The openVA Toolkit for Verbal Autopsies. The R Journal. 2023 Feb;p. 1.

[41] Thomas J, ekarpinskiMITRE, pkmitre, owentrigueros, Choi P, Chu Y.: Pycrossva: Prepare Data from WHO and PHRMC Instruments for Verbal Autopsy Algorithms. Available from: https://pypi.org/project/pycrossva/.