November 1, 2025

Dr. Felix Busch and Diana Corogeanu
Editors, BMC Medicine

Re: Manuscript Revisions

Dear Felix, Diana, and Reviewers

We sincerely thank you for the time and effort dedicated to reviewing our manuscript titled "Computer Assisted Verbal Autopsy: Comparing Large Language Models to Physicians for Assigning Causes to 6939 Deaths in Sierra Leone from 2019–2022". We have revised the entirety of the manuscript to incorporate GPT-5, ensuring its relevance and timeliness, while each reviewer's comment has been addressed in detail with corresponding references to the changes made.

Please kindly find our responses to the reviewers below, with references to tracked changes in the attached document named *wen-et-al-2025-cava-rv1-changes.docx*.

Sincerely,

Professor Prabhat Jha, OC, MD, DPhil
Director, CGHR, University of Toronto
On behalf of Richard Wen and the co-authors

Reviewer 1

I recommend minor revision.

> **Response 1.1:** Thank you, we appreciate the time and effort to provide us with helpful feedback. We are pleased to provide detailed responses to your comments and have referenced each revision relative to the revised manuscript named *wen-et-al-2025-cava-rv1-changes.docx* with tracked changes highlighted in red.

The section on reproducibility should be expanded. The short experiment in Appendix C shows inconsistency of GPT outputs, but the main text could discuss this more explicitly and propose how to ensure stability in practice.

> **Response 1.2:** We have expanded the section on reproducibility. We included clarification in Section 4 that explains the configuration of GPT-3.5/4/5 for more deterministic outputs. We also set a parameter "seed" suggested by OpenAI to provide mostly consistent outputs for GPT-3.5/4 in Appendix B.2.2. Note that provided the examples directly from OpenAI with a recommended temperature of 0 and a constant seed of 1234 for maximizing consistency [1], we still observed that there was variation in the textual output despite the context being similar. Given that GPT models were controlled by an external provider (OpenAI), we also note that parameters were subject to support through OpenAI's Application Programming Interface (API), which varied over newer versions. This was the case when we ran the latest GPT-5 model for this revision, where the previous parameters, "temperature" and "seed" were no longer supported in the updated API for deterministic outputs. Instead, we set "reasoning" and "verbosity" parameters to adjust for more deterministic outputs, as noted in Appendix B.2.2 [2,3]. We also included additional GPT-4/5 sample reruns showing non-deterministic outputs explained in Appendix C. Thus, ensuring complete stability is a difficult endeavour for GPT outputs. Although we agree that ensuring is stability is ideal, to the best of our knowledge, recently available works [4–6], and given our experiment in Appendix C, best efforts to adjust model parameters only currently *increase* stability in practice. We clarify this in Section 4, with references to recent studies on the stability of GPT model outputs demonstrating similar findings.

The discussion on privacy is important, but the reader would benefit from concrete mitigation strategies, for example, the feasibility of local deployment of smaller LLMs or systematic anonymization of narratives before processing.

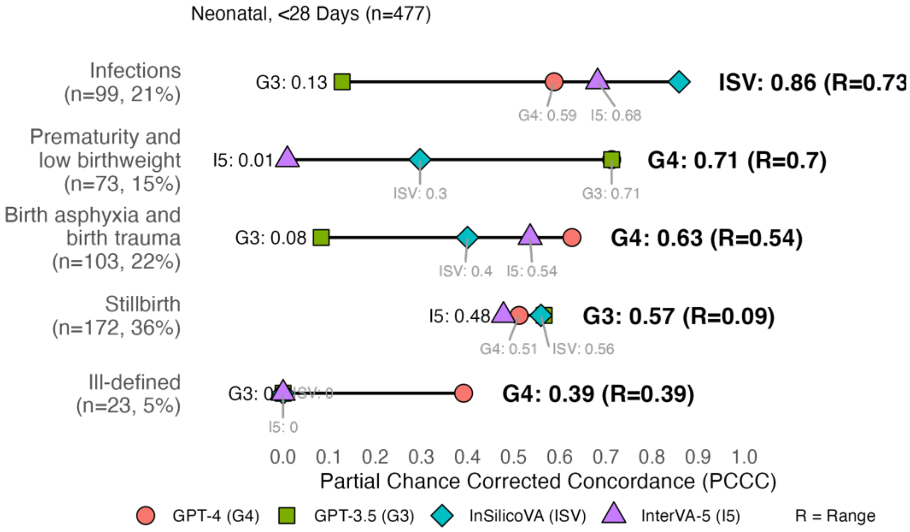
> **Response 1.3:** We have referenced and mentioned data anonymization [7–9] as a common mitigation strategy in Section 4. In the same section, we have also addressed the feasibility of deploying local smaller LLMs by clarifying, with reference to literature [10–12], that it is an available alternative only if adequate expertise and computing resources are met.

Figures 5–7 are very informative, but the dense labeling makes interpretation difficult; simplifying or reformatting them would improve readability.

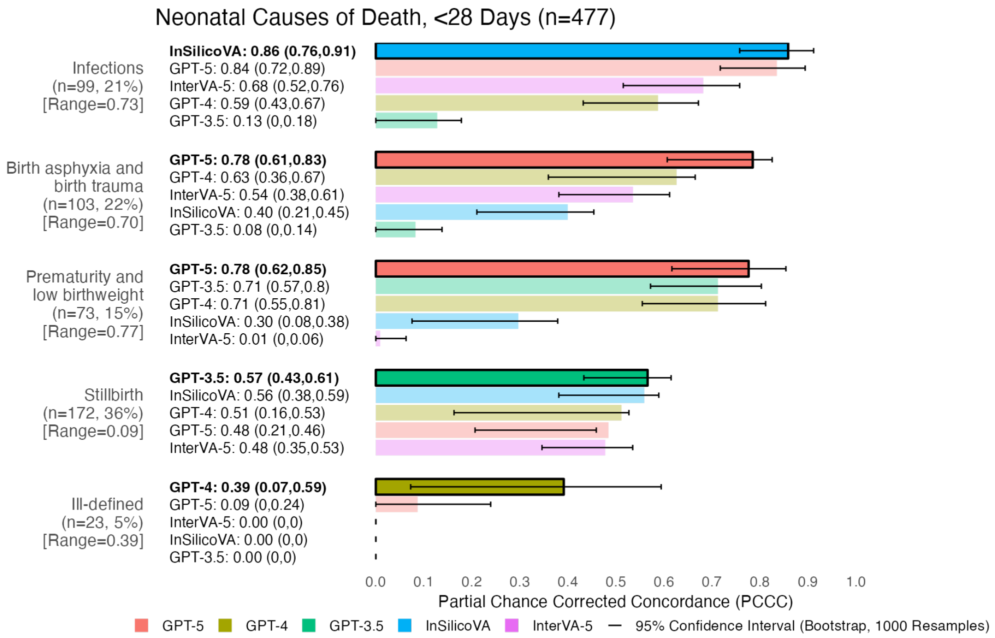
> **Response 1.4:** We have simplified Figure 3 (for consistency) and Figures 5-7 by converting them into grouped bar charts to lower visual crowding of labels for better readability. Error bars and text for the 95% confidence interval were also included for reference. Causes of death and age groups were sorted from the highest to lowest top performing models, while model performance text was grouped vertically into each cause of death and age group category for

improved organization. The reader may conveniently compare performances across categories in top-down fashion, with the familiarity of the standard bar plot. For your convenience, we have provided an example of the neonatal performance plot (Figure 7) before and after to demonstrate this revision.

Before (Figure 7, original manuscript):



After (Figure 7, revised manuscript):



Finally, the generalisability of the results beyond Sierra Leone should be addressed more directly — highlighting to what extent findings may translate to other regions with different mortality profiles.

Response 1.5: We have addressed the generalizability of the results beyond Sierra Leone more directly through mention of regional variations [13–15] and its influence on the

> generalizability of models to other regions. We expect GPT models to outperform statistical models like InSilicoVA and InterVA-5 overall and for most CODs, similar to the results in our paper, as they have leverage up-to-date sources of information from the web (e.g., news, books, manuals, reports) that are often not available or missed in statistical models [16–20]. These sources hold information for generalizing to other regions, such as widely known or changing regional variations in mortality or culture. We also emphasize that these sources are not without bias (e.g., underrepresentation on the web), and performance is not guaranteed [20,21]. Thus, we note the need for further validation studies when applying them to other regions. We have included this revision with references in Section 4.

Reviewer 2

Thank you for giving me the opportunity to review your paper. The experiments are interesting, and the paper is well written. I have a few questions that I would like you to answer and provide additional information.

> **Response 2.1:** Thank you for taking the time to review our manuscript and for providing thoughtful and constructive feedback. We are pleased to offer additional information and clarification. Detailed responses to your comments are provided below with references to each corresponding edit in the revised manuscript named *wen-et-al-2025-cava-rv1-changes.docx*, containing tracked changes highlighted in red.

Issues of methodology

The GPT models are tested on narratives, the methods InterVA-5 and InSilicoVA use the standardised questionnaire, while the physician-assigned COD use both. The authors assume that the questionnaires and the narratives contain the same information. Or that the information on the narrative is reflected by the answers to questions on the questionnaires. Assuming equivalence is risky: even if the narrative seems to contain the same information as the questionnaire, the way this is expressed may be different. Relatives may emphasise different symptoms in the narrative versus when probed by fixed questions (recall bias and interpretation). What is the impact of using different data for the models on the results and comparison?

> **Response 2.2:** We agree that the questionnaire and narratives do not contain equivalent information. We wish to address this confusion and clarify that we do not assume this. Narratives are unstructured text that provide additional context and temporal information (e.g., event circumstances, symptom chronology, behaviours), often missed by the questionnaire, and as mentioned, are expressed differently than from structured responses. In past studies, this information in the narratives has been shown to improve the performance of COD prediction models, in contrast to solely relying on the questionnaire [22–24]. Building on these past studies, we use narratives and questionnaire data (recognizing they do not contain equivalent information) to allow GPT models to leverage the latent patterns in the narratives, while providing comparison with widely adopted models (InterVA-5 [25] and InSilicoVA [26]) that are currently unable to fully utilize narratives. We have included an additional clarification in the methods (Section 2.2). Similar to previous studies, we observed that GPT models utilizing narratives had the impact of consistently outperforming questionnaire-based statistical models (InterVA-5/InSilicoVA) in individual-level performance, as shown in our results (Section 3).

The prompts for the GPT models mentioned the country name. Why was that done and what implications could there be regarding the results returned by the models. Is it possible that the models use information (e.g. country statistics regarding diseases) that they already have, extracted from datasets and academic papers. Have you tried to prompt without giving away the country name?

> **Response 2.3:** Thank you for the insightful comment. When physicians review each Verbal Autopsy (VA) record, the country that the deceased resides in is mostly always given and known. For fair comparison, we wished to also provide GPT models with this standard information, as it is available both to physicians and the widely adopted models InterVA-5 and InSilicoVA. It is possible that GPT models use information, such as country statistics, that they already have as their training data includes academic papers and potentially datasets accessible on the web. However, the black box nature of these models poses a barrier to directly verify if this is the case at a technical level. In addition, we have instructed GPT models to provide ICD-10 codes, for which there are thousands of possibilities, and may potentially be more difficult to apply statistics to. Alternatives include prompting for reasoning output (e.g., asking the GPT model to provide rationale for its assignments or more directly, requesting to know if the country was helpful for its assignment), and, as suggested, a sensitivity analysis to observe its impact on performance by tuning the model (e.g., prompting with and without the country name). As these were not the primary focus of the paper, we have not attempted a thorough analysis on the impact of including the country in the prompt.

Why did the prompt instruct the models to return an ICD-10 code rather than a CGHR-10 Code or the WHO VA cause list? Would the results have been different? Were the ICD-10 codes given to the models or did the models already know them?

> **Response 2.4:** The prompt instructed GPT models to return an ICD-10 code to mirror physician practice. The ICD-10 system is a standard that is widely adopted and well-studied with strict and granular definitions of each code, allowing for comparison across analyses [27–29]. In addition, CGHR-10 and WHO VA codes directly group ICD-10 codes, which provided the flexibility of a standardized coding system while enabling convenient conversion from ICD-10 to CGHR-10, WHO VA, or any other system based on more granular ICD-10 codes. Thus, grouping ICD-10 codes also enables stricter and more standardized definitions of COD categories, ensuring that comparable categories across different coding systems can be accurately aligned, and that observed results are not influenced by variations in coding systems or category definitions [30,31]. We expect that if GPT models would use CGHR-10 codes, all records would be coded instead of the >99% shown in Figure B1, as some ICD-10 codes were not captured by CGHR-10. Consequently, if the models used WHO VA codes, we may see more records lost to conversion. These may slightly impact performance scores depending on the age groups and CODs of the lost records, while mostly negligible for overall performance results. We also suspect that it is possible that performance would improve with the use of CGHR-10 or WHO VA codes instead of ICD-10, as the number of CGHR-10 and WHO VA codes are comparatively less than the thousands of ICD-10 codes. This limits the number of possibilities and thus reduces misclassification [32–35]. However, the caveat is the loss of diagnostic specificity, which may obscure important differences between grouped disease codes (e.g., cancer sites) or lead to concerns related to Response 2.3 above, where GPT models may simply utilize country statistics to assign CODs rather than leveraging the informative patterns in the narratives. Additional rationale has been provided to justify instructing GPT models to use the ICD-10 coding system in Appendix B.2.1. We also included the mention of GPT models outperforming InterVA-5/InSilicoVA models despite the advantage of fewer possible codes to assign from in Section 4. We did not provide ICD-10 codes to the GPT models as they already had knowledge of them. An additional sentence has been added to clarify this detail in Appendix B.2.1.

What do you think the behaviour of the models would have been if the models were asked to return a cause of death in English? Rather than asking the model to "jump straight to the code", asking the models to give both a code and an English COD could have provided a pathway to ICD-10 that could be audited. It may have helped with the issue of unreliability of the GPT-3 answers.

> **Response 2.5:** Thank you for the suggestion. GPT model outputs were based on tokens (units of text such as word segments, phrases, punctuation), and an underlying mechanism that tries to predict the next most likely token. Assuming that reliability refers to the consistency of outputs across repeated runs, the increase of output tokens (e.g., including an English COD) may potentially increase randomness, leading to a reduction in reliability. With the addition of the English COD, more tokens are produced and thus, more possibilities exist due to the flexibility of expressing the same COD with different arrangements of word segments, phrases, and punctuation. This may also introduce issues with standardizing these different arrangements and comparing equivalency in meaning. However, we acknowledge the importance of evaluating the reliability of outputs, and wish to mention that it is an emerging topic for LLMs and approaches are still developing [36]. This mention has been added in the discussion (Section 4).

In the discussion, we are told that GPT-3.5 did not assign consistent CODs when repeated on the same record. This aspect should have been considered as part of the methodology (e.g. design choice to control the randomness of the model, using an ensemble approach, or the suggestion above, prompt engineering). The experiment described that the model gave different answers at different times for the same prompt and input. How about for a 'slightly different' prompt or for a different arrangement in the input data?

> **Response 2.6:** We agree that evaluating GPT models and their reliability is an important consideration for our study. Thus, we have mentioned the reduction of randomness in GPT model outputs, with the objective of producing more deterministic results in Section 2.2. Further details have also been provided in Appendix B.2.2, where we describe each parameter and application of best practices to configure for more deterministic outputs. Despite this, we mention in Section 4, and explain for reviewer 1 in Response 1.2, that a limitation of GPT models is that they are inherently non-deterministic, and configuration of parameters are current best practice, as noted by OpenAI, to *increase* the reliability of outputs. We also note in the same section, that the robust analysis of output reliability for GPT models is out of the scope of this study, scaling substantially with time and costs when hundreds or thousands of repetitions are needed [37–39]. As our study's focus was to compare LLMs leveraging narrative data to widely adopted models that rely on questionnaire responses, the issue of output consistency and attempts to remedy them (e.g., prompt engineering), remain an emerging topic and a limitation in this paper. However, we signify its importance by applying current best practices in parameter selection detailed in Appendix B.2.2, conducting a small sample of experiments in Appendix C, and discussing it along with mention of emerging approaches in Section 4.

Should the performance scores not be adjusted by *an error component*?

> **Response 2.7:** We wish to clarify that the Cause Specific Mortality Fraction (CSMF) accuracy and Partial Chance Corrected Concordance (PCCC) both have error components included in their calculation when measuring model performance. CSMF accuracy measures population-level performance and is adjusted by the maximum total error from the worst model (see Appendix B.3.1), while PCCC measures individual-level performance and is corrected for random chance (see Appendix B.3.2). Both metrics were based on [40]. Our methods were built on well-cited past studies [22,32,41–44] evaluating Computer Coded Verbal Autopsy

(CCVA) models using PCCC and CSMF accuracy as robust metrics. In Section 2.3, we have included clarification of the error adjustments incorporated into each metric, and mentioned that PCCC and CSMF accuracy have both been widely used in assessing performance for CCVA models. In addition, we have computed the 95% confidence intervals using bootstrapping (1000 resamples with replacement) for Figures 3 and 5-7 for reference.

It would be important to know the following: Were the records given to the LLM models to be assessed one by one (in separate sessions?) or as a block? Was there a possibility of "memory contamination"?

> **Response 2.8:** Thank you for bringing this to our attention. For consistency across GPT models, the records given to the LLM models (GPT-3.5/4/5) were assessed one by one in separate sessions (also known as stateless API calls). Thus, we suspect that "memory contamination" in this manner and context is likely not of concern, to the best of our knowledge. We have included this information in Appendix B.2.2, detailing the APIs we had used and our configuration for statelessness, in addition to describing the setting of the *store* parameter to enable statelessness for GPT-5. Note that GPT-3.5/4 were stateless by default, and did not require any parameters for statelessness.

Was the data in some particular order or contained any groupings, e.g. all children tested together or in blocks?

> **Response 2.9:** The data was in the order of record entry by surveyors and was tested in two blocks representing a survey round each. Each block contained all adult (12 to 69 years), child (28 days to 11 years), and neonate (under 28 days) records collected for the survey round. The data was tested in this manner as we received and processed the verbal autopsy data as survey rounds were completed.

Some other observations

Figure B1 – suggests that questionnaires were also input to GPT and that narratives were input to InterVA. Is that the case?

> **Response 2.10:** Thank you for bringing our attention to this detail. We would like to clarify that only the age and sex were taken from the questionnaire and used to construct the prompts for GPT models. We have revised Figure B1 in Appendix B.1 to address this confusion.

Should Table 1 from discussion be moved to results?

> **Response 2.11**: We have moved Table 1 and its accompanying sentence from the discussion to the results in Section 3.1.

Table B4 Study data by cause of death: how does the agreement between physicians for different conditions compare to that of the models.

> **Response 2.12**: We have included agreement data in Tables B4 and B5 for reference. We highlighted notable CODs with high and low physician agreement in Section B.1. Additionally, the performance of models in relation to physician agreement for CODs was added to the results in Sections 3.2 to 3.4. For the discussion in Section 4, we noted physician agreement in our data and the high performance of the models for most CODs, which provides the opportunity for improving physician agreement with computer assisted verbal autopsy.

References

1. OpenAI, Anadkat S. How to make your completions outputs consistent with the new seed parameter [Internet]. OpenAI Cookbook. 2023 [cited 2025 Oct 9]. Available from: https://cookbook.openai.com/examples/reproducible_outputs_with_the_seed_parameter
2. OpenAI. Using GPT-5 [Internet]. OpenAI API. 2025 [cited 2025 Oct 14]. Available from: https://platform.openai.com/docs/guides/latest-model
3. OpenAI, Singh M. GPT-5 New Params and Tools [Internet]. OpenAI Cookbook. 2025 [cited 2025 Oct 14]. Available from: https://cookbook.openai.com/examples/gpt-5/gpt-5_new_params_and_tools
4. Krishna S, Bhambra N, Bleakney R, Bhayana R, Atzen S. Evaluation of Reliability, Repeatability, Robustness, and Confidence of GPT-3.5 and GPT-4 on a Radiology Board–style Examination. Moy L, editor. Radiology. 2024 May 1;311(2):e232715.
5. Jang ME, Lukasiewicz T. Consistency Analysis of ChatGPT [Internet]. arXiv; 2023 [cited 2024 July 6]. Available from: http://arxiv.org/abs/2303.06273
6. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq. 2023 Feb 28;rs.3.rs-2566942.
7. Olatunji IE, Rauch J, Katzensteiner M, Khosla M. A Review of Anonymization for Healthcare Data. Big Data. 2024 Dec 1;12(6):538–55.
8. Vovk O, Piho G, Ross P. Methods and tools for healthcare data anonymization: a literature review. International Journal of General Systems. 2023 Apr 3;52(3):326–42.
9. Zuo Z, Watson M, Budgen D, Hall R, Kennelly C, Al Moubayed N. Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study. JMIR Med Inform. 2021 Oct 15;9(10):e29871.
10. Das BC, Amini MH, Wu Y. Security and Privacy Challenges of Large Language Models: A Survey. ACM Comput Surv. 2025 June 30;57(6):1–39.
11. Wang F, Lin M, Ma Y, Liu H, He Q, Tang X, et al. A Survey on Small Language Models in the Era of Large Language Models: Architecture, Capabilities, and Trustworthiness. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V2 [Internet]. Toronto ON Canada: ACM; 2025 [cited 2025 Aug 10]. p. 6173–83. Available from: https://dl.acm.org/doi/10.1145/3711896.3736563
12. Corradini F, Leonesi M, Piangerelli M. State of the Art and Future Directions of Small Language Models: A Systematic Review. Big Data and Cognitive Computing. 2025;9(7):189.
13. Setel PW, Macfarlane SB, Szreter S, Mikkelsen L, Jha P, Stout S, et al. A scandal of invisibility: making everyone count by counting everyone. The Lancet. 2007;370(9598):1569–77.
14. Byass P. The Imperfect World of Global Health Estimates. PLoS Med. 2010 Nov 30;7(11):e1001006.
15. Fottrell E, Byass P. Verbal autopsy: methods in transition. Epidemiologic reviews. 2010;32(1):38–55.
16. Wang S, Zhu Y, Liu H, Zheng Z, Chen C, Li J. Knowledge Editing for Large Language Models: A Survey. ACM Comput Surv. 2024 Nov 11;57(3):59:1-59:37.
17. Nalmpatian A, Heumann C, Alkaya L, Jackson W. Transfer learning for mortality risk: A case study on the United Kingdom. PLoS One. 2025;20(5):e0313378.
18. Singh H, Mhasawade V, Chunara R. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. PLOS Digital Health. 2022 Apr 5;1(4):e0000023.
19. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023;619(7969):357–62.
20. Zhang Z, Zhao J, Zhang Q, Gui T, Huang X. Unveiling Linguistic Regions in Large Language Models [Internet]. arXiv; 2024 [cited 2025 Oct 20]. Available from: http://arxiv.org/abs/2402.14700

21. Dunn J, Adams B, Madabushi HT. Pre-Trained Language Models Represent Some Geographic Populations Better Than Others [Internet]. arXiv; 2024 [cited 2025 Oct 20]. Available from: http://arxiv.org/abs/2403.11025
22. Jeblee S, Gomes M, Jha P, Rudzicz F, Hirst G. Automatically determining cause of death from verbal autopsy narratives. BMC Med Inform Decis Mak. 2019 July 9;19(1):127.
23. 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):1315–25.
24. 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 Med Res Methodol. 2016 Feb 1;16(1):13.
25. 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 30;17(1):102.
26. McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark SJ. Probabilistic Cause-of-death Assignment using Verbal Autopsies. J Am Stat Assoc. 2016;111(515):1036–49.
27. Henderson T, Shepheard J, Sundararajan V. Quality of diagnosis and procedure coding in ICD-10 administrative data. Medical care. 2006;44(11):1011–9.
28. Mezzich JE. International surveys on the use of ICD-10 and related diagnostic systems. Psychopathology. 2002;35(2–3):72–5.
29. Hirsch JA, Nicola G, McGinty G, Liu RW, Barr RM, Chittle MD, et al. ICD-10: history and context. American Journal of Neuroradiology. 2016;37(4):596–9.
30. Watzlaf VJ, Garvin JH, Moeini S, Anania-Firouzan P. The effectiveness of ICD-10-CM in capturing public health diseases. Perspectives in Health Information Management/AHIMA, American Health Information Management Association. 2007;4:6.
31. Hsu MC, Wang CC, Huang LY, Lin CY, Lin FJ, Toh S. Effect of ICD-9-CM to ICD-10-CM coding system transition on identification of common conditions: An interrupted time series analysis. Pharmacoepidemiology and Drug Safety. 2021;30(12):1653–74.
32. 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 4;12(1):20.
33. Ramirez-Villalobos D, Stewart AL, Romero M, Gomez S, Flaxman AD, Hernandez B. Analysis of causes of death using verbal autopsies and vital registration in Hidalgo, Mexico. PLOS ONE. 2019 July 3;14(7):e0218438.
34. King G, Lu Y. Verbal autopsy methods with multiple causes of death. 2008 [cited 2025 Oct 23]; Available from: https://projecteuclid.org/journals/statistical-science/volume-23/issue-1/Verbal-Autopsy-Methods-with-Multiple-Causes-of-Death/10.1214/07-STS247.short
35. Murray CJ, Lopez AD, Black R, Ahuja R, Ali SM, Baqui A, et al. Population Health Metrics Research Consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. Population Health Metrics. 2011 Aug 4;9(1):27.
36. Liu X, Chen T, Da L, Chen C, Lin Z, Wei H. Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V2 [Internet]. Toronto ON Canada: ACM; 2025 [cited 2025 Oct 23]. p. 6107–17. Available from: https://dl.acm.org/doi/10.1145/3711896.3736569
37. Mosbach M, Andriushchenko M, Klakow D. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines [Internet]. arXiv; 2021 [cited 2025 Oct 23]. Available from: http://arxiv.org/abs/2006.04884
38. Dodge J, Ilharco G, Schwartz R, Farhadi A, Hajishirzi H, Smith N. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping [Internet]. arXiv; 2020 [cited 2025 Oct 23]. Available from: http://arxiv.org/abs/2002.06305

39. Byrne MD. How many times should a stochastic model be run? An approach based on confidence intervals. In: Proceedings of the 12th International conference on cognitive modeling, Ottawa [Internet]. 2013 [cited 2025 Oct 23]. Available from: https://iccm-conference.neocities.org/2013/proceedings/papers/0083/paper0083.pdf
40. Murray CJ, Lozano R, Flaxman AD, Vahdatpour A, Lopez AD. Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. Population Health Metrics. 2011 Aug 4;9(1):28.
41. 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 June 27;17(1):116.
42. 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 Med. 2014 Feb 4;12(1):22.
43. Murtaza SS, Kolpak P, Bener A, Jha P. Automated verbal autopsy classification: using one-against-all ensemble method and Naïve Bayes classifier. Gates open research. 2019;2:63.
44. Miasnikof P, Giannakeas V, Gomes M, Aleksandrowicz L, Shestopaloff AY, Alam D, et al. Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. BMC Med. 2015 Dec;13(1):286.