Computer Assisted Verbal Autopsy: Comparing Large Language Models to Physicians for Assigning Causes to 6939 Deaths in Sierra Leone from 2019-2022

Richard Wen[1], Anteneh Tesfaye Assalif[1,2], Andy Sze-Heng Lee[1], Rajeev Kamadod[1], Asha Behdinan[1], Ronald Carshon-Marsh[1], Catherine Meh[1], Thomas Kai Sze Ng[1], Patrick Brown[1], Prabhat Jha[1*], Rashid Ansumana[2]

[1*]Centre for Global Health Research, St. Michael's Hospital, Unity Health Toronto and University of Toronto, 30 Bond St, Toronto, M5B 1W8, Ontario, Canada.
[2]School of Community Health Sciences, Njala University, Bo, Sierra Leone.

*Corresponding author(s). E-mail(s): prabhat.jha@utoronto.ca;
Contributing authors: rwen@torontomu.ca; antenehta@gmail.com; andylee.dee@gmail.com; rajeevk@kentropy.com; asha.behdinan@mail.utoronto.ca; ronald.carshonmarsh@mail.utoronto.ca; catherine.meh@unityhealth.to; kaisze.ng@unityhealth.to; patrick.brown@utoronto.ca; prabhat.jha@utoronto.ca; rashidansumana@gmail.com;

Abstract

Background: Verbal autopsies (VAs) collect information on deaths occurring outside traditional healthcare settings to estimate representative Causes of Death (CODs). Current computer models assign CODs at population-level accuracy comparable to physicians, but perform poorly at the individual-level, largely due to reliance on structured questionnaire data and neglect of narrative free text. Recently, the large language model ChatGPT-4 demonstrated human-level performance on professional and academic benchmarks. While ChatGPT-4 shows promise in COD assignment, its application to VA narratives has not yet been evaluated.
Methods: We analyzed 6,939 VA records from Sierra Leone (2019–2022) to compare four models, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA, against physician-assigned CODs at population and individual levels. GPT models used narratives, whereas InterVA-5 and InSilicoVA relied on questionnaires. CODs were grouped into 19, 10, and 7 categories for adult, child, and neonatal deaths. Cause Specific Mortality Fraction (CSMF) accuracy and Partial Chance Corrected Concordance (PCCC) were used to assess population and individual-level agreement with physician coding respectively, stratified by age and COD.
Results: GPT-4 outperformed all models overall (PCCC=0.61), followed by GPT-3.5 (0.56) then InSilicoVA and InterVA-5 (0.44). GPT-4 achieved the highest performance for adult (0.64) and neonatal deaths (0.58), while GPT-3.5 had the highest performance for child deaths (0.54). Across ages, performance increased from 1 month to 14 years and declined from 15 to 69 years. GPT4, GPT-3.5, and InSilicoVA achieved the highest PCCC in 17, 9, and 4 of the 30 CODs, respectively. At the population level, all models achieved comparable CSMF accuracies (0.74–0.79).
Conclusion: All models performed similarly at the population level, but GPT models and InSilicoVA showed greater performance for specific CODs at the individual-level. GPT models demonstrated improvements over InterVA-5 and InSilicoVA models. This study provides foundational evidence for integrating computer models to assist physicians with alternative diagnoses, helping reduce ill-defined codes and improve agreement in COD assignment.

Keywords: Cause of Death, Physicians, Computer-Assisted Diagnosis, Artificial Intelligence, Natural Language Processing, Machine Learning, Mortality, Surveillance, Mathematical Models, Global Health

1 Background

Every year, 41 million people died prematurely from noncommunicable diseases, accounting for 74% of all deaths globally [1]. While most of these deaths are preventable, effective intervention requires evidence-based resource allocation that targets high-risk populations [2]. Reliable mortality counts and accurate Cause of Death (COD) data are essential for guiding public health policy and reducing premature mortality [3–6]. However, civil registration and vital statistics systems remain incomplete in many low-income countries. Fewer than half of all deaths are registered, and among these, only 8% have an assigned COD [7]. To address this gap, Verbal Autopsy (VA) has been deployed as a scalable method for collecting mortality data and assigning likely CODs, particularly for deaths that occur outside of healthcare facilities, which account for more than half of all deaths [8–11].

VA involves two major components: survey and COD assignment [12–14]. In the survey component, trained interviewers use structured questionnaires and open narrative prompts to gather data from relatives or close contacts of the deceased. In the COD assignment component, physicians review these data to determine the most likely COD. However, reliance on physician assignment has been criticized for limited reproducibility and subjectivity [15–19]. To overcome these limitations, automated Computer Coded Verbal Autopsy (CCVA) methods such as InterVA [20] and InSilicoVA [17] have been developed. These models offer scalable and reproducible alternatives and have demonstrated comparable performance to physicians at the population level. However, their performance at the individual-level remains limited [21–25], while their reliance on structured questionnaire data often omits open narrative text, which can contain additional contextual and chronological information that may improve diagnostic accuracy [26–28].

Recent advances in large language models (LLMs), trained on vast textual datasets using deep learning methods, have significantly improved natural language processing (NLP) capabilities. These include tasks such as question answering, code generation, and medical reasoning based on free text [29–32]. ChatGPT, developed by OpenAI and released in 2022, is a widely accessible LLM capable of generating human-like responses to natural language queries. Earlier versions (GPT-1 to GPT-3) scaled from 117 million to 175 billion parameters and were trained on data ranging from 5 GB to 45 TB [33]. In 2023, ChatGPT-4 was introduced, achieving human-level performance on a range of academic and professional benchmarks [34].

Given the underutilization of narrative free text in VA analysis and the capabilities of LLMs in processing

such data, we conducted a study using VA records from Sierra Leone (2019–2022) to compare four models, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA, against physician-assigned CODs. This work aims to evaluate the potential of LLMs in enhancing COD assignment from narrative data in low-resource settings.

2 Methods

This study outlines the methodology used to compare cause of death (COD) assignments from four models, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA, with physician-determined CODs, as summarized in Figure 1. The dataset was first filtered to include only records with physician agreement, as described in Section 2.1. Section 2.2 details the input formats and output structures of the four models. Section 2.3 presents the evaluation framework, which compares model outputs to physician assigned CODs using both population-level and individual-level performance metrics. Additional details are provided in Appendix B.

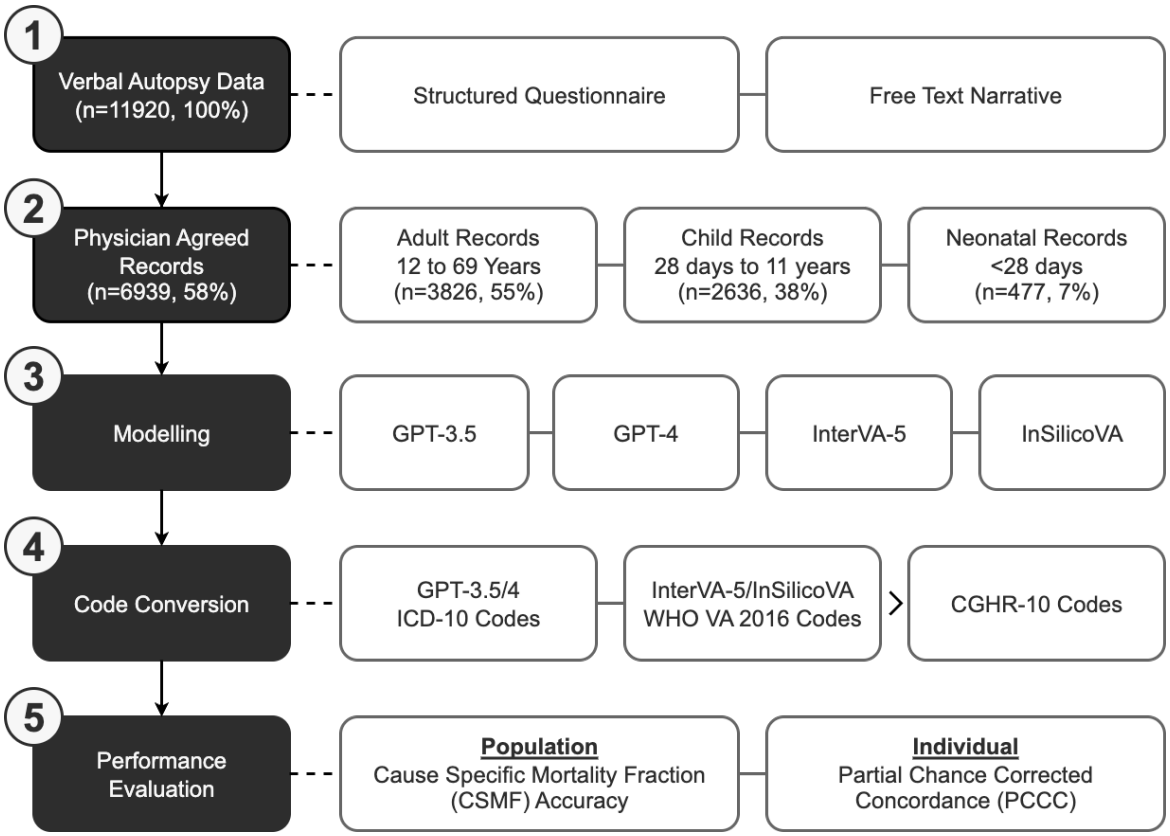

Fig. 1 Flow diagram for verbal autopsy coding comparison of 6939 sample deaths in Sierra Leone. Verbal autopsy data containing 11,920 sample deaths were initially collected from in-field surveys, and filtered to 6939 records where two randomly assigned physicians agreed on the cause of death. Four computer models GPT-3.5, GPT-4, InterVA-5, and InSilicoVA were compared to physicians using standardized CGHR-10 codes, and evaluated using individual PCCC and population CSMF accuracy metrics.

2.1 Verbal Autopsy (VA) Data

A total of 11,920 verbal autopsy (VA) records were obtained from the HEAL-SL study [35, 36], which employed dual-coded Electronic Verbal Autopsy (EVA). Each record was independently reviewed by two randomly selected physicians, who assigned COD codes based on the International Classification of Diseases, 10th Revision (ICD-10) [37]. Agreement between physician-assigned CODs was evaluated using Central Medical Evaluation Agreement 10 (CMEA-10) codes, which group related ICD-10 codes into broader, clinically similar categories [38] (see Additional File 1). If both codes fell within the same CMEA-10 group, the record was considered in agreement. Disagreements entered a reconciliation phase, where each physician was shown both the assigned codes and the reasoning from the other physician. Physicians could then (1) retain their original code, (2) adopt the other physician's code, or (3) assign a new code. Records that remained unresolved proceeded to adjudication, where a senior physician reviewed all reasoning and assignments and issued a final COD.

To ensure comparability with physician coding, only records with physician agreement were used in this study, as such cases provide higher confidence in the COD assignment [18, 39, 40]. From the original dataset, 6,942 records met this criterion. All ICD-10 codes were then standardized to CGHR-10 categories (see Appendix A), which group causes into 19, 10, and 7 categories for adults (12–69 years), children (28 days to 11 years), and neonates (under 28 days), respectively. After excluding three records without a valid CGHR-10 category, a total of 6,939 physician-agreed records (3,826 adult, 2,636 child, and 477 neonatal) were used for model comparison and performance evaluation. Further details on data preprocessing are provided in Appendix B.1, with COD and age group distributions summarized in Tables B4 and B5.

2.2 Modelling

Four computational models were used to assign causes of death (CODs) for each of the 6,939 physician-agreed verbal autopsy (VA) records: GPT-3.5, GPT-4, InterVA-5, and InSilicoVA. InterVA-5 and InSilicoVA are widely used statistical models within the OpenVA framework for COD assignment in VAs [13, 21, 22, 24, 25, 41–43]. InterVA-5 applies a Bayesian probabilistic approach, using a standardized set of symptoms and expert-derived conditional probabilities to assign the most likely COD based on maximum probability [20, 44, 45]. InSilicoVA extends this approach by incorporating a hierarchical Bayesian framework and Markov

Chain Monte Carlo (MCMC) methods [46–48], allowing for quantification of uncertainty, individual-level probability estimates, and the integration of additional data sources [17]. GPT-3.5 [49] and GPT-4 [34] are large language models (LLMs) that generate human-like text by learning from large amounts of textual data [50]. These models are trained using reinforcement learning from human feedback [51–54], enabling them to follow natural language instructions and generate human-level responses. GPT-4 demonstrated improvements over GPT-3.5, including more recent training data, enhanced reasoning capabilities, and multimodal input-output functionality (e.g. text, image, voice) [33].

For GPT-3.5 and GPT-4, the following user prompt was used to instruct each model to produce COD assignments as ICD-10 codes, where *<age>* and *<sex>* from the questionnaire, and *<narrative>* from the narratives, were replaced with values from the data:

Determine the underlying cause of death and provide the most probable ICD–10 code for a verbal autopsy narrative of a <age> years old <sex> death in Sierra Leone: <narrative>

InterVA-5 and InSilicoVA used structured questionnaire data, which were converted into OpenVA-compatible format [43]. Both models produced COD assignments coded using the WHO 2016 VA standard [55]. To ensure comparability across models, all output CODs were mapped to the CGHR-10 classification system for evaluation relative to physician-assigned CODs. Further details on model input formats, output mappings, and code conversion procedures are provided in Appendix B.2.

2.3 Performance Evaluation

Model performance was assessed at both the population and individual levels by comparing each model's CGHR-10 COD assignments to those of physicians for all 6,939 records. Cause-Specific Mortality Fraction (CSMF) accuracy was used to evaluate agreement at the population level (see Appendix B.3.1), while Partial Chance Corrected Concordance (PCCC) was used to assess individual-level agreement (see Appendix B.3.2) [56]. Both metrics range from 0 to 1, where higher values indicate stronger similarity with physician assignment. Given that model performance can vary by age and different CODs [41, 42, 57], both CSMF accuracy and PCCC were calculated overall and stratified by age group (adult, child, neonatal), CGHR-10 COD, and age at death. For adult and child groups, metrics were computed in five-year age bands for records with age at death of one year or older, and five-month bands for records between 28 days and one year. For

the neonatal group, evaluations were conducted separately for age intervals of 0–6 days and 7–27 days. Additional details on the evaluation strategy and metric calculations are provided in Appendix B.3.

3 Results

3.1 Overall Performance

Population level performances were similar for all models (0.74-0.79 CSMF accuracy). Thus, the remainder of the results focus on the individual-level performances measured by PCCC. GPT-4 demonstrated the highest performance (0.61) followed by GPT-3.5 (0.56), InSilicoVA (0.44), and InterVA-5 (0.44) (Figure 2). GPT-3.5 and GPT-4 had improvements from 0.14-0.18 in the PCCC over InSilicoVA and InterVA-5, while GPT-4 slightly improved over GPT-3.5 by 0.05. Figure 3 shows the individual performance across three age groups (adult, child, and neonate). GPT-4 had the best performance for adult and neonatal records (0.64 and 0.58), while GPT-3.5 had the best performance for child records (0.54) with GPT-4 performing slightly worse (0.51). InSilicoVA and InterVA-5 performed the worst for adult and child records (less than 0.5), while GPT-3.5 performed the worse for neonatal records (0.42). Performance varied less for child deaths (range: 0.13) than for adult and neonatal deaths (range: 0.24 and 0.22). Across ages, all models followed a similar pattern in performance (Figure 4). In adults, performance decreased with age for 12 to 59 years (from 0.7 down to 0.35), suggesting greater difficulty in assigning CODs among older adults, with a modest improvement observed after age 59. Among children and neonates, performance improved from 5 months to 11 years (from 0.1 towards 0.75), indicating greater model reliability as developmental age advanced. The highest and lowest performances were observed for ages 12-29 years (0.4-0.7) and 1-11 months (0.1-0.35) respectively.

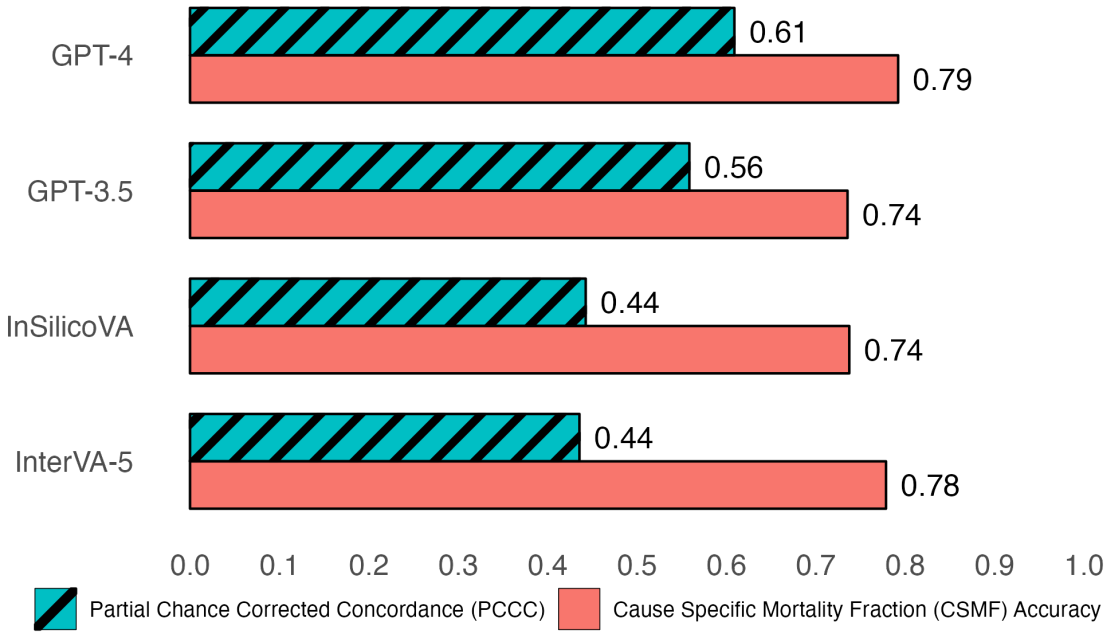

Fig. 2 Individual (PCCC) and population (CSMF accuracy) level verbal autopsy coding performance of all 6939 deaths. PCCC and CSMF accuracy values range from 0 to 1. PCCC values of 1 indicate complete agreement with physician coding per individual death, while CSMF accuracy values of 1 indicate complete agreement with physician coding per cause, irrespective of individual deaths.

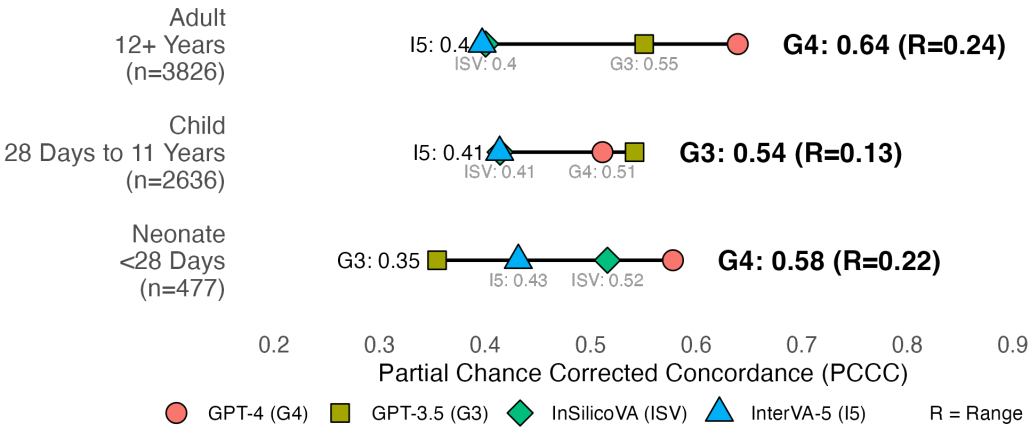

Fig. 3 Individual-level verbal autopsy coding performance by age group. PCCC values range from 0 to 1, with 1 indicating complete agreement with physician coding per individual death. R (range) represents the difference between the maximum and minimum PCCC values across all models per age group.

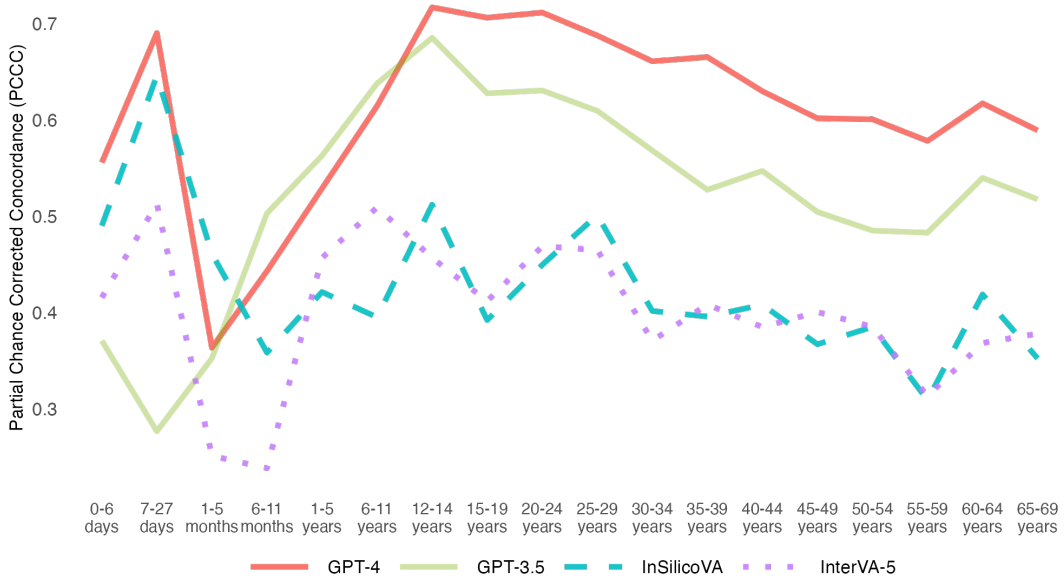

Fig. 4 Individual-level model performance by age of deceased. PCCC values range from 0 to 1, with 1 indicating complete agreement with physician coding per individual death. Ages 0-27 days represent neonatal deaths, ages 1-11 months represent child deaths, and ages 12-69 years represent adult deaths.

3.2 Performance for 3826 Adult Records (12 to 69 years)

Figure 5 presents performance across 17 adult CODs. GPT-4 achieved the highest individual-level performance for 10 of 17 CODs (0.35–0.99), followed by GPT-3.5 for 5 CODs (0.43–0.94), and InSilicoVA for 2 CODs (0.71 and 0.84). InterVA-5 had the lowest performance for 8 CODs (0–0.79), InSilicoVA for 6 CODs (0.01–0.41), and GPT-3.5 for 2 CODs (0.38 and 0.53). The greatest improvements of GPT-3.5/4 over InSilicoVA and InterVA-5 were observed in chronic respiratory diseases (0.74-0.94 in the PCCC), while the smallest improvements were for malaria (0.09-0.17 in the PCCC). All models performed well for maternal conditions (0.79–0.99), but poorly for unspecified infections (0.35–0.49), malaria (0.26–0.43), and ill-defined CODs (0–0.35). GPT-4 showed performance improvements between over all other models for cancers (0.25-0.36 in the PCCC), stroke (0.27–0.45 in the PCCC), and diarrhoeal diseases (0.37–0.51 in the PCCC). GPT-3.5 demonstrated similar gains for liver and alcohol-related diseases (0.27–0.52 in the PCCC). Performance variability across models was most pronounced for chronic respiratory diseases (range: 0.94), while narrower differences were observed for maternal conditions (range: 0.20), malaria (range: 0.17), ischemic heart disease (range: 0.15), and unspecified infections (range: 0.14).

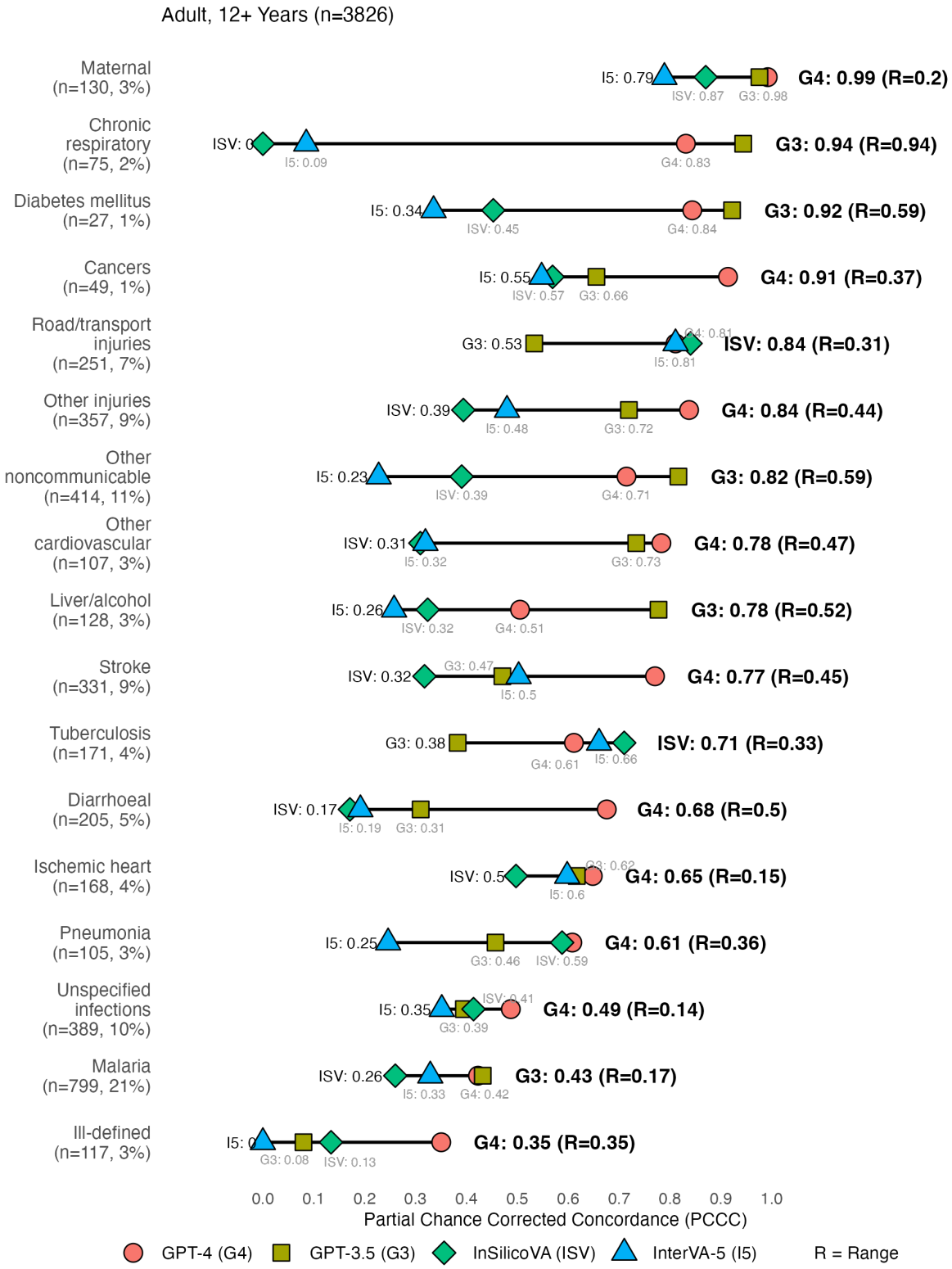

Fig. 5 Individual-level model performance for adult causes of death. PCCC values range from 0 to 1, with 1 indicating complete agreement with physician coding per individual death. R (range) represents the difference between the maximum and minimum PCCC values across all models per cause of death. Symbols on the far left represent lowest performing models, while symbols on the right with bolded text represent highest performing models per cause of death. Suicide (n=3, <1%) was excluded due to low sample size.

3.3 Performance for 2636 Child Records (28 Days to 11 Years)

Figure 6 shows individual-level performance across 8 child CODs, excluding congenital anomalies due to a low sample size (n=1, <1%). GPT-4 achieved the highest PCCC for 4 of the 8 CODs (0.65–0.94), followed by GPT-3.5 for 3 CODs (0.44–0.88), and InSilicoVA for 1 COD (0.78). InterVA-5 had the lowest performance for 4 CODs (0.09–0.79), InSilicoVA for 3 CODs (0–0.35), and GPT-3.5 for 1 COD (0.58). All models performed well for injuries (0.79–0.94), while showing lower performance for malaria (0.35–0.54) and other infections (0.29–0.44). GPT-4 demonstrated an improvement over other models for ill-defined CODs, with improvements between 0.38–0.65 in the PCCC, while demonstrating stronger performance for injuries, with gains of 0.11-0.15 compared to 0.01-0.04 in the PCCC for other models. Performance differences exceeding 0.60 in the PCCC were observed for epilepsy, leukaemia, other communicable diseases (range: 0.73), ill-defined causes (range: 0.65), and nutritional deficiencies (range: 0.61). In contrast, narrower differences (less than 0.30 in the PCCC) were seen for malaria (range: 0.20), injuries, and other infections (range: 0.15).

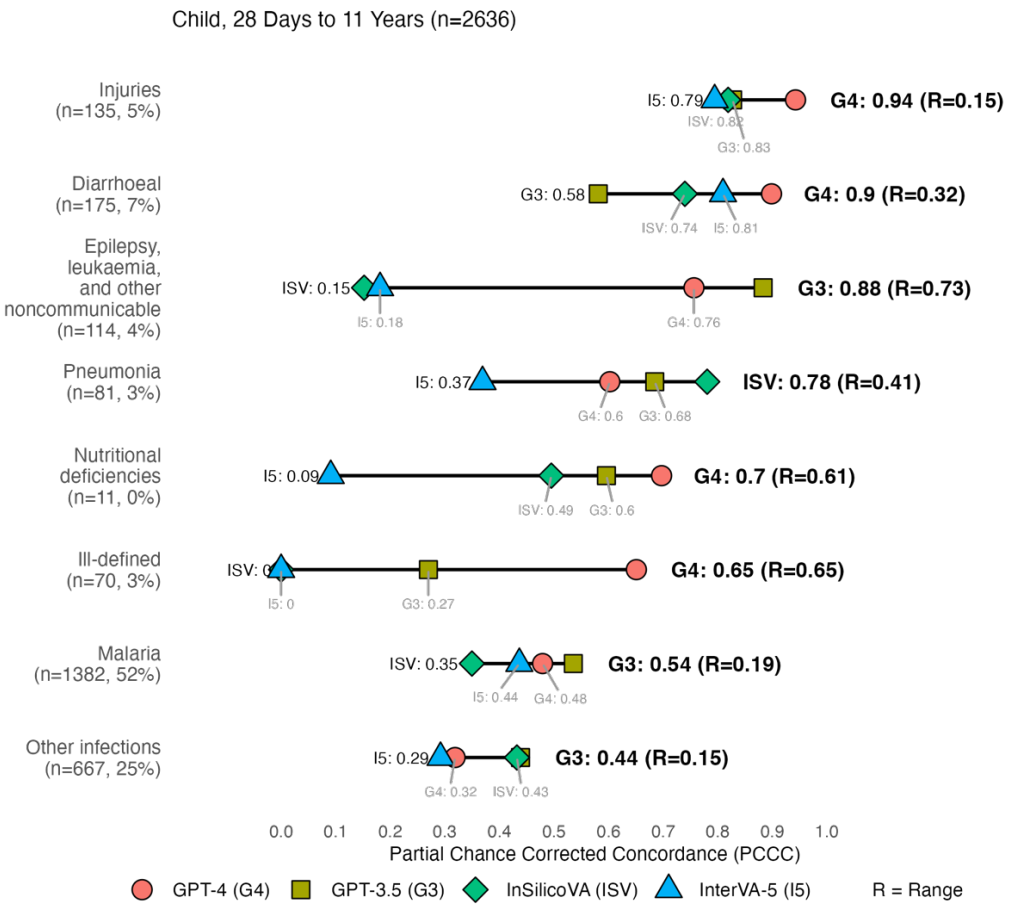

Fig. 6 Individual-level model performance for child causes of death. PCCC values range from 0 to 1, with 1 indicating complete agreement with physician coding per individual death. R (range) represents the difference between the maximum and minimum PCCC values across all models per cause of death. Symbols on the far left represent lowest performing models, while symbols on the right with bolded text represent highest performing models per cause of death. Congenital anomalies (n=1, <1%) was excluded due to low sample size.

3.4 Performance for 477 Neonatal Records (Under 28 Days)

Figure 7 shows model performance across 5 neonatal CODs, excluding congenital anomalies (n=2, <1%) and other causes (n=5, 1%) due to limited sample sizes. GPT-4 achieved the highest performance for 3 of the 5 CODs (0.39–0.71), while GPT-3.5 and InSilicoVA had the highest performance for one COD each (0.57 and 0.86). GPT-3.5 showed the lowest performance for 3 CODs (0–0.13), and InterVA-5 for 2 CODs (0.01 and 0.48). Performance was similar across all models for stillbirths (0.48–0.57). Notably, only GPT-4 achieved a PCCC greater than zero for prematurity-related deaths. InSilicoVA outperformed all other models for neonatal infections, with gains of 0.18–0.73 in the PCCC. Larger performance differences between models were observed for infections (range: 0.73) and prematurity and low birthweight (0.7), while lower differences were seen in stillbirth (range: 0.09).

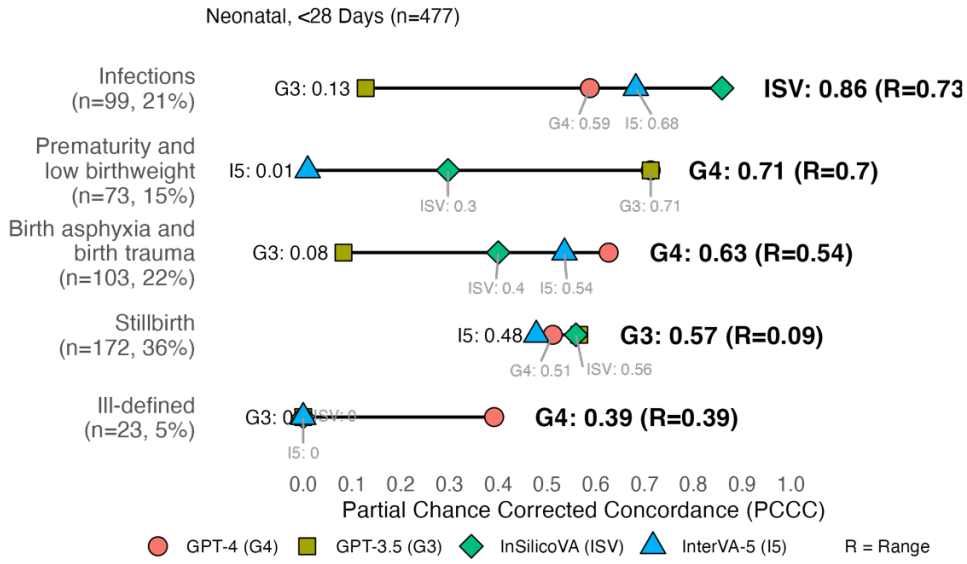

Fig. 7 Individual-level model performance for neonatal causes of death. PCCC values range from 0 to 1, with 1 indicating complete agreement with physician coding per individual death. R (range) represents the difference between the maximum and minimum PCCC values across all models per cause of death. Symbols on the far left represent lowest performing models, while symbols on the right with bolded text represent highest performing models per cause of death. Congenital anomalies (n=2, <1%) and other causes (n=5, 1%) were excluded due to low sample size.

4 Discussion

As model performance varied by disease and age, the findings suggest cause-specific models to maximize performance across disease categories, and ensuring that performance across age align with expectations from clinical literature as validation [58, 59]. In terms of individual performance, GPT-3.5/4 consistently outperformed InterVA-5 and InSilicoVA for most leading CODs identified in prior Sierra Leone studies [36, 60] as seen in Table 1. A key advantage of GPT-3.5/4 is their ability to process and generate natural language text as input and output. Unlike InterVA-5 and InSilicoVA, GPT models assign CODs using the ICD-10 standard, mirroring physician practice. In contrast, InterVA-5 and InSilicoVA rely exclusively on structured WHO VA 2016 questionnaires and assign CODs using broader WHO VA 2016 codes. This dependency necessitates ongoing maintenance and conversion between questionnaire versions and coding systems, reducing interoperability and comparability across models. In addition, rarer diseases, underrepresented in questionnaire data, are better contextualized through external knowledge (e.g., web sources, journals, books) embedded in GPT models. The flexibility of GPT models in handling unstructured data allows them to capture latent and ambiguous information, such as health-seeking behaviors and social context, which are not encompassed by standardized VA codes or structured questionnaires [26, 28].

Although GPT models improved over InterVA-5 and InSilicoVA models, several limitations exist. A brief experiment in Appendix C revealed that GPT-3.5 did not assign consistent CODs when repeated on the same record [61–63]. In contrast, InterVA-5 and InSilicoVA provide assignments with probabilities for alternative causes, which was made feasible by calculating probabilities using repeated runs without costs. Another limitation common to all models was their reliance on past training data, limiting potential to detect new or emerging diseases (e.g., COVID-19). This is often remedied with re-training or updating models with new data or knowledge [64–66]. We also note that GPT-3.5/4 required data sent to external servers, raising significant privacy concerns from reliance on third-party services [67, 68]. While jurisdictions, such as the European Union, enforce strict protections under the General Data Protection Regulation (GDPR), most low- and middle-income countries are only beginning to formalize regulatory frameworks for data protection and artificial intelligence governance [69–71]. In contrast, InterVA-5 and InSilicoVA are run on local systems under the control of the data owner. As technology improves, larger GPT models may be possible on local systems, while currently, smaller LLMs exist as an alternative [72–74]. Although this study rigorously compares computer algorithms for COD assignment in Sierra Leone, the extent to which these findings are generalizable in other geographic or epidemiological contexts remains limited. Given ongoing efforts to scale and integrate VA systems for mortality surveillance across diverse low- and middle-income countries, further validation across globally representative VA datasets is essential to evaluate model robustness, adaptability, and operational utility in practice [75–77].

This study establishes a basis for Computer Assisted Verbal Autopsy (CAVA), the integration of computer models into VA systems to support physician assignment. Model-generated COD suggestions can be offered to physicians after their initial assignment, enabling reconsideration or confirmation of CODs as seen in step 2 of Figure 8. We highlight that the integration of models into VAs is both scalable and affordable in resource-constrained settings [39]. At the time of analysis, GPT-3.5 cost ~$0.02 USD per 100 records, GPT-4 cost ~$1.65 USD per 100 records [78], while InterVA-5 and InSilicoVA were freely available as open-source software. These costs complement physician review, which is already affordable at ~$1 USD per household (including field survey) in settings like India [15, 16]. Given recent studies supporting improvement in physician diagnosis from LLM assistance [79, 80], we foresee the potential of alternative COD suggestions from computer models reducing physician disagreement and frequency of ill-defined

records. Presently, we have integrated CAVA (using GPT-4, InterVA-5, and InSilicoVA) into the ongoing HEAL-SL study [35], with future work evaluating the impact of CAVA on physician assignment.

Table 1 Top ten leading causes of death for Sierra Leone in 2023 and best performing models for verbal autopsy coding.

Top 10 Leading Cause of Death (~71% of ~76K deaths)[1]	Deaths (% of 76K)[2]	Best Model(s) at the individual-level
Malaria	16,075 (21%)	GPT-3.5/4
Infections	11,777 (16%)	GPT-3.5/4/InSilicoVA
Ischaemic heart and other vascular	5,747 (8%)	GPT-4
Diarrhoea	4,285 (6%)	GPT-4
Stroke	4,262 (6%)	GPT-4
Pneumonia	3,074 (4%)	GPT-4/InSilicoVA
Birth asphyxia and birth trauma	2,431 (3%)	GPT-4
Tuberculosis	2,399 (3%)	InSilicoVA
Low birth weight/preterm	1,570 (2%)	GPT-4
Asthma and chronic respiratory	1,551 (2%)	GPT-3

[1]Other infections and severe systemic/localized infections were generalized into infections. Appendix, hernia, intestinal and Peptic ulcer/gastroesophageal causes did not have comparable CGHR-10 codes and were omitted from the top ten.
[2]Percentage of ~76 thousand (K) total deaths [60]. Numbers are rounded.

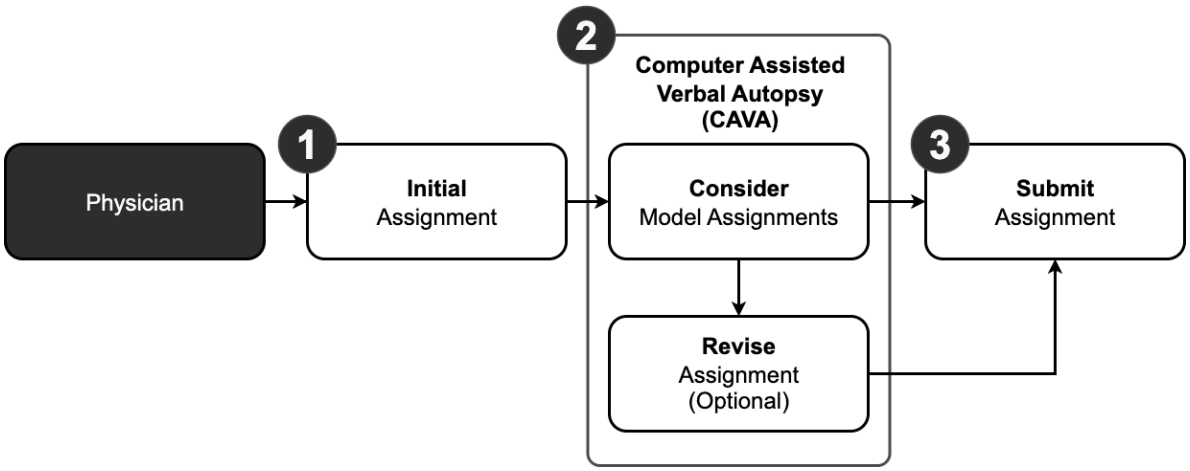

Fig. 8 Computer Assisted Verbal Autopsy (CAVA) integrated into physician coding as a three-step process. The first step involves the physician assigning an initial cause of death to a record without considering causes of death provided by models. The second step is the addition of CAVA, where the physician can compare their initial assignment in step one to model assignments and optionally choose to revise their initial assignments. The third step submits the record with either the initial or revised assignment to the verbal autopsy coding system.

5 Conclusion

This study evaluated the performance of GPT-3.5, GPT-4, InterVA-5, and InSilicoVA models against physicians in assigning CODs for 6,939 VA records from Sierra Leone (2019–2022). At the population level, all models achieved similar CSMF accuracy (0.74–0.79). At the individual-level, GPT-4 had the highest performance (0.61 PCCC), followed by GPT-3.5 (0.58), and InSilicoVA/InterVA-5 (0.44). By COD, GPT-4 performed best for 10 of 17 adult, 4 of 8 child, and 3 of 5 neonatal causes, while GPT-3.5 led in 5 adult, 3

child, and 1 neonatal CODs, and InSilicoVA led in 2 adult, 1 child, and 1 neonatal cause. Performance increased (∼0.1–0.75 PCCC) as children and neonates matured (0 days to 14 years) and decreased (∼0.7–0.35) with adult aging (15 to 69 years). These findings suggest that combining models tailored to specific CODs and age groups may optimize performance relative to physicians. All models demonstrated scalability and on-demand availability, enabling COD estimation and alternative diagnoses in low-resource or physician-scarce settings. GPT models' natural language processing capability allowed flexible data input and output, aligning closer to physician reasoning, but issues remain with reproducibility, reliance on historical training data, computational demands, and data privacy. Study limitations included challenges comparing ICD-10 codes across models, limited sensitivity analyses due to costs, and exclusion of multiple COD assignment evaluation. Future research opportunities include prompt engineering and custom GPT models to improve accuracy, guided household surveys to enhance narrative quality, and CAVA systems integrating GPT and other models to support physicians by suggesting alternative COD assignments. GPT-4, InterVA-5, and InSilicoVA have been incorporated into ongoing HEAL-SL study since 2022 to provide second-opinion support for physician COD assignment. Evaluating the impact of computer-assisted VA on physician agreement and reduction of ill-defined deaths will be critical to advancing accurate, efficient VA systems worldwide.

Supplementary information. Additional file 1 (.csv) titled "Central Medical Evaluation Agreement 10 (CMEA-10) codes" with description" ICD-10 code ranges considered in physician agreement" was used to supplement this study.

Acknowledgments. TBD.

Declarations

Ethics approval and consent to participate

The Sierra Leone Ethics and Scientific Review Committee (SLESRC No. 025/04/2023) and Unity Health Toronto Research Ethics Board (REB#15-231) granted ethics approval for the project. Relatives of the deceased provided informed consent.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the Open Mortality repository, https://openmortality.org, on reasonable request. All code files used and/or analysed during the current study are available in the Github repository, https://github.com/cghr-toronto/healsl-gpt-paper.

Competing interests

The authors had no conflicts of interest.

Funding

Bill & Melinda Gates Foundation, Canada Institutes of Health Research. The sponsors had no role in the design, implementation, data collection, analyses, or report preparation.

Authors' contributions

PJ and RA are the study Principal Investigators. ATA and RK implemented the data collection procedures. RW, TKSN, and CM processed, documented, and prepared the data. RW, ASL, and RK ran the models. RW wrote the paper and conducted the analysis. AB and RCM provided medical domain guidance and feedback. All authors reviewed the results and contributed to the report. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Appendix A CGHR-10 Codes

Table A1 CGHR-10 codes for adults (12-69 years).

Adult CGHR-10 Code	ICD-10 Range
Acute respiratory infections	H65-H68, H70-H71, J00-J22, J32, J36, J85-J86, P23
Tuberculosis	A15-A16, B90, J65
Diarrhoeal	A00-A09
Unspecified infections	A17-A33, A35-A99, B00-B17, B19-B49, B55-B89, B91-B99, C46, D64, D84, G00-G09, H10, H60, I30, I32-I33, K02, K04-K05, K61, K65, K67, K81, L00-L04, L08, M00-M01, M60, M86, N10, N30, N34, N41, N49, N61, N70-N74, P35-P39, R50, R75, ZZ21
Malaria	B50-B54
Maternal conditions	A34, F53, O00-O08, O10-O16, O20-O99
Nutritional deficiencies	D50-D53, E00-E02, E40-E46, E50-E64, X53-X54
Chronic respiratory	J30-J31, J33-J35, J37-J64, J66-J84, J90-J99, R04-R06, R84, R91
Cancers	C00-C26, C30-C45, C47-C58, C60-C97, D00-D48, D91, N60, N62-N64, N87, R59
Ischemic heart	I20-I25, R55
Stroke	G45-G46, G81-G83, I60-I69
Diabetes mellitus	E10-E14
Other cardiovascular	I00-I03, I05-I15, I26-I28, I31, I34-I52, I70-I99, R00-R01, R03, ZZ23
Liver and alcohol related	B18, F10, K70-K77, R16-R18, X45, Y15, Y90-91
Other noncommunicable	D55-D63, D65-D83, D86, D89, E03-E07, E15-E35, E65-E68, E70E90, F00-F09, F11-F52, F54-F99, G10-G37, G40-G41, G43-G44, G50-G80, G84-G99, H00-H06, H11-H59, H61-H62, H69, H72-H95, K00-K01, K03, K06-K14, K20-K31, K35-K38, K40-K60, K62-K64, K66, K78-K80, K82-K93, L05, L10-L99, M02-M54, M61-M85, M87-M99, N00-N08, N11-N29, N31-N33, N35-N40, N42-N48, N50N59, N75-N86, N88-N99, Q00-Q99, R10-R15, R19-R23, R26-R27, R29-R49, R56, R63, R70-R74, R76-R77, R80-R82, R85-R87, R90, ZZ25
Road and transport injuries	V01-V99, Y85
Suicide	X60-X84
Other injuries	S00-S99, T00-T99, W00-W99, X00-X44, X46-X52, X55-X59, X85-X99, Y00-Y14, Y16-Y84, Y86, Y89, Y92-Y98, ZZ27
Ill-defined	R02, R07-R09, R25, R51-R54, R57-R58, R60-R62, R64-R69, R78-R79, R83, R89, R92-R94, R96, R98-R99

334 **Table A2** CGHR-10 codes for children and neonates.

Child CGHR-10 Code (28 days to 11 years)	ICD-10 Range
Pneumonia	A37, H65-H68, H70-H71, J00-J22, J32, J36, J85-J86, P23, U04
Diarrhoeal	A00-A09
Malaria	B50-B54
Other infections	A15-A28, A30-A36, A38-A44, A46, A48-A71, A74-A75, A77-A99, B00B09, B15-B27, B30, B33-B49, B55-B60, B64-B83, B85-B92, B94-B97, B99, G00-G09, H10, H60, I30, I32-I33, I39-I41, J65, K02, K04-K05, K61, K65, K67, K81, L00-L04, L08, M00-M01, M60, M86, N10, N30, N34, N41, N49, N61, N70-N74, P35-P39, R50, R75, U00, Y95, ZZ11
Congenital anomalies	P01, P05, P07, P21, Q00-Q99
Epilepsy, leukaemia, and other noncommunicable	C00-C97, D01-D48, D55-D89, E03-E35, E65-E90, F00-F02, F73, G10G99, H00-H06, H11-H59, H61-H62, H69, H72-H95, I00-I28, I31, I34-I38, I42-I99, J30-J31, J33-J35, J37-J47, J60, J64, J66-J70, J80-J82, J84, J90-J99, K00-K01, K03, K06-K60, K62-K63, K70-K80, K82-K93, L05, L10-L99, M02-M54, M61-M85, M87-M99, N00-N08, N11-N29, N31N33, N35-N40, N42-N48, N50-N51, N60, N62-N64, N75-N99, P04, P08, P27, P51, P53-P60, P70-P72, P74-P76, P78, P80-P83, P92-P94, R00-R01, R03-R06, R11-R23, R26-R27, R29-R49, R55-R56, R59, R63, R70-R74, R76-R77, R80-R82, R84-R87, R90-R91, ZZ12-ZZ13, ZZ15
Injuries	S00-S99, T00-T98, V01-V99, W00-W99, X00-X52, X57-X99, Y00-Y91, Y97-Y98
Nutritional deficiencies	D50-D53, E00-E02, E40-E46, E50-E56, E59-E61, E63-E64, X53-X54
Other	D00, F03-F72, F74-F99, P00, P02-P03, P10-P15, P20, P22, P24-P26, P28-P29, P50, P52, P61, P77, P90-P91
Ill-defined	P96, R02, R07, R09-R10, R25, R51-R54, R57-R58, R60-R62, R64, R68-R69, R78-R79, R83, R89, R92-R99
Neonate CGHR-10 Code (<28 days)	
Prematurity & low birthweight	D64, O60, P01, P05, P07, P22, P25-P28, P52, P61, P77, P80, P92, R04
Neonatal infections	A00-A09, A20-A28, A32-A35, A37-A44, A46, A48-A49, A68A70, A74-A75, A77-A79, A81-A90, B54, B95-B96, G00-G09, H10, H60, H65-H68, H70-H71, I30, I32-I33, I39-I41, J00-J22, J32, J36, J85-J86, K65, K67, K81, L00-L04, L08, M00-M01, M60, M86, N10, N30, N34, N41, N49, N61, O85, P23, P35,P39, P58-P59, P63, U04
Birth asphyxia and birth trauma	G40, P00, P02-P03, P10-P15, P20-P21, P24, P29, P50-P51, P90-P91, R06, W79, Z37
Stillbirth	P95
Congenital anomalies	C76, Q00-Q99
Other	A15-A19, A30-A31, A36, A50-A67, A71, A80, A91-A99, B00-B09, B15-B27, B30, B33-B53, B55-B60, B64-B83, B85-B92, B94, B97, B99, C00-C75, C77-C97, D00-D48, D50-D53, D55-D63, D65-D89, E00-E35, E40-E46, E50-E56, E59-E61, E63-E90, F00-F99, G10-G39, G41-G99, H00-H06, H11-H59, H61-H62, H69, H72-H95, I00-I28, I31, I34-I38, I42-I99, J30, J31, J33-J35, J37-J47, J60, J64-J70, J80-J82, J84, J90J99, K00-K63, K70-K80, K82-K93, L05, L10-L99, M02-M54, M61-M85, M87-M99, N00-N08, N11-N29, N31-N33, N35N40, N42-N48, N50-N51, N60, N62-N64, N70-N99, P04, P08, P53-P57, P60,P70-P72, P74-P76, P78, P81-P83, P93-P94, R00-R01, R03-R05, R11-R23, R26-R27, R29-R36, R39-R50, R55-R56, R59, R63, R70-R77, R80-R82, R84-R87, R90-R91, S00-S99, T00-T98, U00, V01-V99, W00-W78, W80-W99, X00-X54, X57-X99, Y00-Y91, Y95, Y97-Y98
Ill-defined	P96, R02, R07, R09-R10, R25, R51-R54, R57-R58, R60-R62, R64, R68-R69, R78-R79, R83, R89, R92-R99

335

Appendix B Details on Methods

This section provides additional details on the methods described in Section 2. An overview of the methods used in this study is seen in Figure B1 as a five-step process. Section B.1 provides details on the preprocessed data used for modelling. Section B.2 describes the data and parameter inputs and outputs for each model, while Section B.3 details the evaluation of model outputs at the individual and population level across different CODs, age groups, and ages.

B.1 CGHR-10 Physician Agreed Records

Initially, 11,920 records were collected from dual-coded EVA in the HEAL-SL study. 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). 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 (based on comparisons using CMEA-10 codes, see Additional File 1). The 6942 records were converted into CGHR-10 codes (see Appendix A) that generalized ICD-10 codes into 19, 10, and 7 categories for the adult (12 to 69 years), child (28 days to 11 years), and neonatal (under 28 days) age groups. After conversion, a final total of 6939 physician agreed records (3826 adult, 2636 child, and 477 neonatal) were used for modelling and performance evaluation, where three records were removed as their ICD-10 codes did not have a matching CGHR-10 code.

The 6939 physician agreed records were collected using VA from the HEAL-SL study between 2019-2022, where records were collected using nation-wide samples across Sierra Leone provinces seen in Figure B2. More populous areas (e.g. southern and north east provinces with ~197,000 and ~135,000 population respectively) had more sampling areas versus less populous areas (e.g. north west and eastern provinces with ~50,000 and ~69,000 people respectively). The distribution of the study data are shown by CGHR-10 causes of death in Table B4. All age groups had relatively evenly distributed female and male records (44-55% of 6939 records each). Across CODs, there were noticeably more female records for cancers (65%), and maternal conditions (100%), while more male records for chronic respiratory diseases (61%), other

362 noncommunicable diseases (61%), other injuries (77%), road and transport injuries (71%), and
363 tuberculosis (68%). Most records were coded by physicians as malaria for adults (20%) and children
364 (52%), and stillbirth (36%) and neonatal infections (21%) for neonates. Suicide, congenital anomalies,
365 nutritional deficiencies, and other had low sample sizes for each age group (<1% of total records for each
366 age group). Table B5 shows the distribution of the study data by age. Across ages, there were more male
367 records for 50-59 years (60-62%), while all other records had between 49-59% female and male records.
368 Most records were in the 65-69 years age range for adults (15%), 1-5 years for children (62%), and 0-6
369 days for neonates (83%).

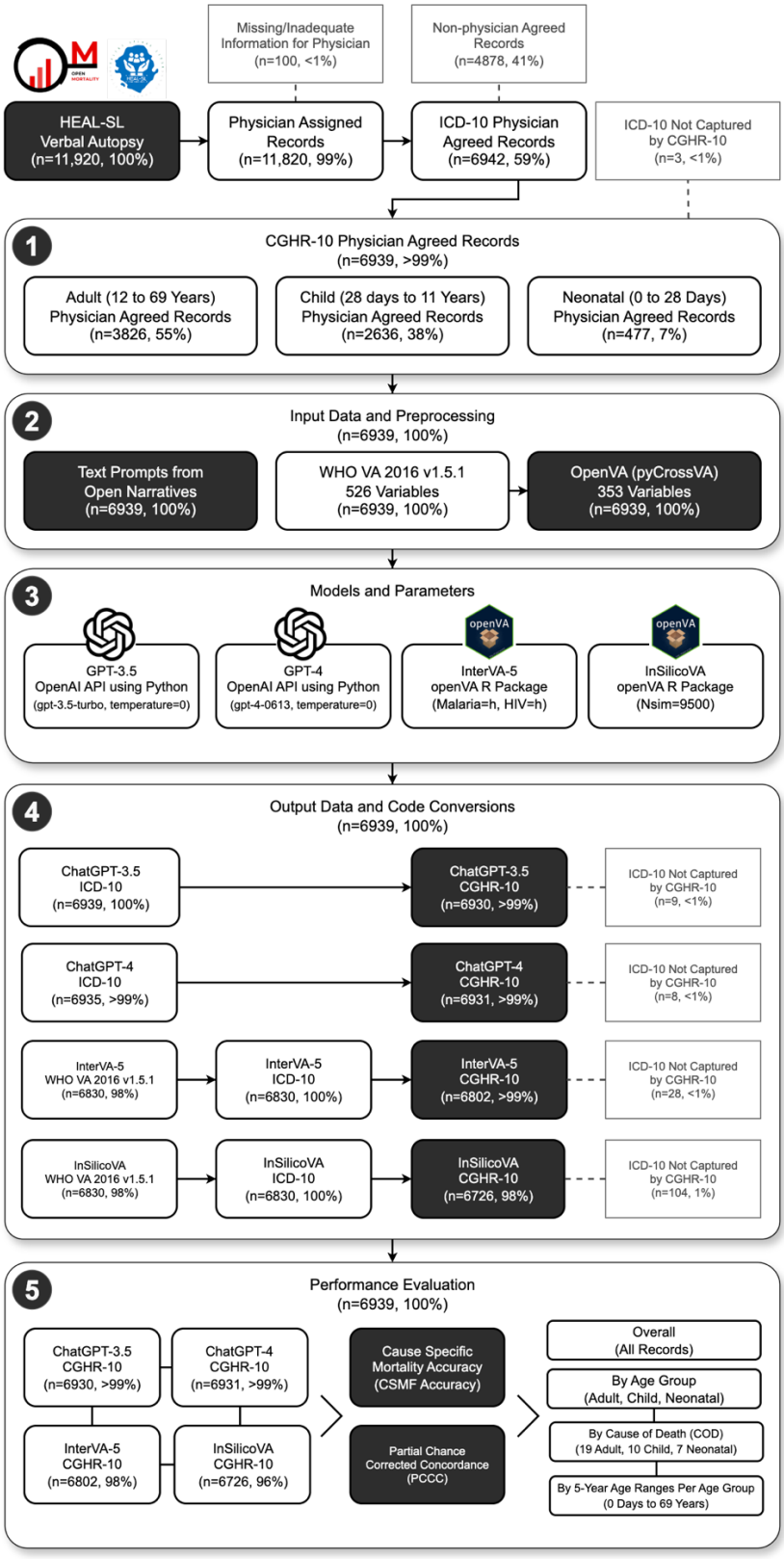

Fig. B1 Detailed flow diagram for verbal autopsy coding comparison of 6939 sample deaths in Sierra Leone. This supplements the diagram in Figure 1, providing additional details on removed records, model specifications, and sub-processes.

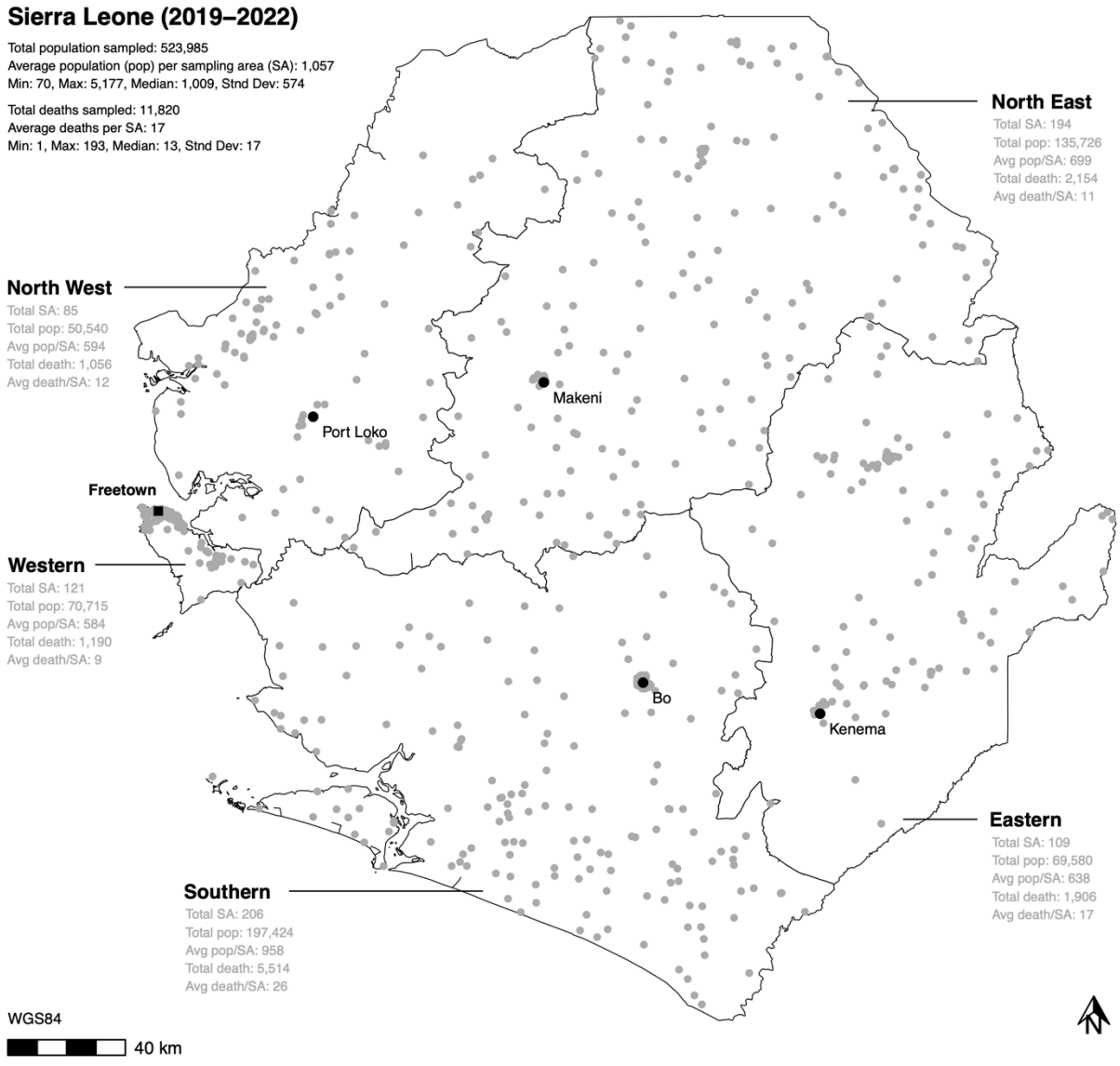

Fig. B2 Healthy Sierra Leone (HEAL-SL) verbal autopsy data sampling areas for Sierra Leone from 2019-2022. The boundaries represent Sierra Leone provinces. Each grey point represents the centroid of sampling areas. Pop=population, SA=sampling area.

B.2 Modelling Details

Each model (GPT-3.5, GPT-4, InSilicoVA, and InterVA-5) required pre-processing of the 6939 records into input data, and standardization of output COD codes from models for performance evaluation as not all models produced comparable codes across outputs. Although each model can assign multiple CODs per record, only the first generated COD response from GPT-3.5 and GPT-4, and the most probable COD from InterVA-5 and InSilicoVA were used for evaluation. Section B.2.1 describes the input data and parameters for each model, while Section B.2.3 details the outputs from running each model.

Table B4 Study data by cause of death.

Age Group	CGHR-10 Cause of Death (COD)	Female	Male	Total
Adult, 18 CODs	Acute Respiratory Infections	48 (45.7%)	57 (54.3%)	105 (2.7%)
(n=3826, 55.1%)	Cancers	32 (65.3%)	17 (34.7%)	49 (1.3%)
Adult Female	Chronic Respiratory Diseases	29 (38.7%)	46 (61.3%)	75 (2%)
(n=1681, 43.9%)	Diabetes Mellitus	14 (51.9%)	13 (48.1%)	27 (0.7%)
Adult Male	Diarrhoeal Diseases	102 (49.8%)	103 (50.2%)	205 (5.4%)
(n=2145, 56.1%)	Ill-Defined	56 (47.9%)	61 (52.1%)	117 (3.1%)
	Ischemic Heart Disease	89 (53%)	79 (47%)	168 (4.4%)
	Liver And Alcohol Related Diseases	58 (45.3%)	70 (54.7%)	128 (3.3%)
	Malaria	372 (46.6%)	427 (53.4%)	799 (20.9%)
	Maternal Conditions	130 (100%)	N/A	130 (3.4%)
	Other Cardiovascular Diseases	59 (55.1%)	48 (44.9%)	107 (2.8%)
	Other Noncommunicable Diseases	160 (38.6%)	254 (61.4%)	414 (10.8%)
	Other Injuries	83 (23.2%)	274 (76.8%)	357 (9.3%)
	Road And Transport Injuries	73 (29.1%)	178 (70.9%)	251 (6.6%)
	Stroke	147 (44.4%)	184 (55.6%)	331 (8.7%)
	Suicide	N/A	3 (100%)	3 (0.1%)
	Tuberculosis	54 (31.6%)	117 (68.4%)	171 (4.5%)
Child, 9 CODs	Congenital Anomalies	1 (100%)	N/A	1 (0%)
(n=2636, 38%)	Diarrhoeal Diseases	79 (45.1%)	96 (54.9%)	175 (6.6%)
Child Female	Epilepsy, Leukaemia, & Other Non-communicable	61 (53.5%)	53 (46.5%)	114 (4.3%)
(n=1290, 48.9%)	Ill-Defined	34 (48.6%)	36 (51.4%)	70 (2.7%)
Child Male	Injuries	51 (37.8%)	84 (62.2%)	135 (5.1%)
(n=1346, 51.1%)	Malaria	680 (49.2%)	702 (50.8%)	1382 (52.4%)
	Nutritional Deficiencies	7 (63.6%)	4 (36.4%)	11 (0.4%)
	Other infections	338 (50.7%)	329 (49.3%)	667 (25.3%)
	Pneumonia	39 (48.1%)	42 (51.9%)	81 (3.1%)
Neonate, 7 CODs	Birth Asphyxia and Birth Trauma	38 (36.9%)	65 (63.1%)	103 (21.6%)
(n=477, 6.9%)	Congenital Anomalies	2 (100%)	N/A	2 (0.4%)
Neonate Female	Ill-Defined	11 (47.8%)	12 (52.2%)	23 (4.8%)
(n=227, 47.6%)	Neonatal Infections	49 (49.5%)	50 (50.5%)	99 (20.8%)
Neonate Male	Other	2 (40%)	3 (60%)	5 (1%)
(n=250, 52.4%)	Prematurity And Low Birthweight	39 (53.4%)	34 (46.6%)	73 (15.3%)
	Stillbirth	86 (50%)	86 (50%)	172 (36.1%)

B.2.1 Input Data and Preprocessing

For GPT-3.5 and GPT-4, 6939 text prompts were generated for each physician agreed record as input to instruct the models to assign CODs based on the open narratives. Two types of text prompts were used: user prompts and system prompts. System prompts contained textual instructions to assign the role of a physician ICD-10 coder with expertise in Sierra Leone. The following system prompt was used:

You are a physician with expertise in determining underlying causes of death in Sierra Leone by assigning the most probable ICD−10 code for each death using verbal autopsy narratives. Return only the ICD−10 code without description. E. g. A00. If there are multiple ICD−10 codes, show one code per line.

User prompts contained textual instructions to perform coding of VA records based on the age, sex, and narrative of the deceased. The following template was used to generate user prompts for each record, where <age> and <sex> from the questionnaire, and <narrative> from the narratives, were replaced with values from the data:

Determine the underlying cause of death and provide the most probable ICD−10 code for a verbal autopsy narrative of a <age> years old <sex> death in Sierra Leone: <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 [81, 82] consisting of 526 variables [83], followed by further conversion into OpenVA format [43] consisting of 353 variables [84] using the pyCrossVA version 0.97 Python package [85]. The 6939 records were all converted into OpenVA formatted records for InterVA-5 and InSilicoVA.

Table B5 Study data by age range.

Age Group	Age Range	Female	Male	Total
Adult (n=3826, 55.1%)	12-14 Years	51 (37.8%)	84 (62.2%)	135 (3.5%)
Adult Female (n=1681, 43.9%)	15-19 Years	115 (42.8%)	154 (57.2%)	269 (7%)
Adult Male (n=2145, 56.1%)	20-24 Years	146 (53.1%)	129 (46.9%)	275 (7.2%)
	25-29 Years	159 (45.2%)	193 (54.8%)	352 (9.2%)
	30-34 Years	174 (50.9%)	168 (49.1%)	342 (8.9%)
	35-39 Years	153 (45.4%)	184 (54.6%)	337 (8.8%)
	40-44 Years	134 (42%)	185 (58%)	319 (8.3%)
	45-49 Years	148 (47%)	167 (53%)	315 (8.2%)
	50-54 Years	134 (39.6%)	204 (60.4%)	338 (8.8%)
	55-59 Years	96 (37.6%)	159 (62.4%)	255 (6.7%)
	60-64 Years	128 (40.8%)	186 (59.2%)	314 (8.2%)
	65-69 Years	243 (42.3%)	332 (57.7%)	575 (15%)
Child (n=2636, 38%)	1-5 Months	146 (47.4%)	162 (52.6%)	308 (11.7%)
Child Female (n=1290, 48.9%)	6-11 Months	160 (50.8%)	155 (49.2%)	315 (11.9%)
Child Male (n=1346, 51.1%)	1-5 Years	822 (50.3%)	811 (49.7%)	1633 (61.9%)
	6-11 Years	162 (42.6%)	218 (57.4%)	380 (14.4%)
Neonate (n=477, 6.9%)	0-6 Days	184 (46.6%)	211 (53.4%)	395 (82.8%)
Neonate Female (n=227, 47.6%)	7-27 Days	43 (52.4%)	39 (47.6%)	82 (17.2%)
Neonate Male (n=250, 52.4%)				

B.2.2 Models and Parameters

The GPT-3.5 and GPT-4 Application Programming Interface (API) was accessed using Python version 3.11.4 and used to assign CODs for each record. GPT-3.5 used the *gpt-3.5-turbo* model, while GPT-4 used the *gpt-4-0613* model. The parameter *temperature* for GPT-3.5 and GPT-4, representing the sampling temperature ranging from 0 to 2 (default of 1), was set to 0 to produce more deterministic outputs [86]. Higher values closer to 2 may produce less deterministic outputs, while lower values closer to 0 produce more deterministic outputs.

The *openVA* R package was used to run InterVA-5 and InSilicoVA models to assign CODs for each record in R version 4.3.1. The *openVA* package version 1.1.1 used dependent packages InterVA5 version 1.1.3 and InSilicoVA version 1.4.0. The *Nsim* (number of iterations to run) parameter [87] for InSilicoVA was set to

9500, while the HIV (level of prevalence of human immunodeficiency virus) and Malaria (level of prevalence of Malaria) parameters [88] for InterVA-5 were both set to *h* (high) reflecting HIV and Malaria disease assumptions in Sierra Leone [89, 90]. Note that the default value of *Nsim=10000* for InSilicoVA ran until 9500 iterations before it stopped due to errors, thus *Nsim=9500* was used and ran successfully.

B.2.3 Output Data and Code Conversion

Of the 6939 input records, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA were able to assign CODs for 6939 (100%), 6935 (>99%), 6830 (98%), 6830 (98%) records respectively. All 6830 (100%) InterVA-5 and InSilicoVA records with WHO VA 2016 v1.5 output codes [55] were converted into ICD-10 codes respectively. After all model outputs were converted to ICD-10 codes, they were further converted to CGHR-10 codes. The 6939 GPT-3.5 and 6935 GPT-4 output records with ICD-10 codes were converted into 6930 (>99%) and 6931 (>99) records with CGHR-10 codes, where <1% (9 and 8) records did not have matching CGHR-10 codes respectively. The 6830 InterVA-5 and InSilicoVA records with ICD-10 codes were converted into 6802 (>99%) and 6726 (98%) records with CGHR-10 codes respectively, where 28 (<1%) and 104 (1%) of records could not be converted into CGHR-10 codes.

B.3 Performance Evaluation Details

The performance of GPT-3.5, GPT-4, InSilicoVA, and InterVA-5 models were evaluated with metrics at the population and individual-level by comparing their CGHR-10 COD outputs for 6939 records to physician COD assignments. Section B.3.1 describes CSMF accuracy in detail for evaluating models on the population level, Section B.3.2 describes PCCC for evaluating models on the individual-level. Records that were assigned a COD by physicians, but not by a model were considered an incorrect COD assignment by the model. CSMF accuracy and PCCC were calculated for each model overall and by three age groups (adult, child, and neonatal), then further into age and COD for each age group.

B.3.1 Cause Specific Mortality Fraction (CSMF) Accuracy

CSMF accuracy measures the performance of models at the population level, comparing distributions of CODs between the physicians and the models [56]. To calculate CSMF accuracy, $CSMF_j$ was calculated as is the fraction of physician or model records for cause *j*, given by dividing the number of records for cause *j*

with the total number of records as seen in Equation B1. Then, the *CSMFMaximumError*, representing the worst possible model, is calculated using Equation B2. Finally, the CSMF accuracy is given by Equation B3, where k is the number of causes, j is a cause, $CSMF_j^{true}$ is the true physician CSMF for cause j, and $CSMF_j^{pred}$ is the prediction model CSMF for cause j. CSMF accuracy ranges from 0 to 1, where 1 means that the model completely matched the physician COD distribution and 0 means that it did not match the distribution.

$$CSMF_j = \frac{Records_j}{Records} \tag{B1}$$

$$CSMFMaximumError = 2(1 - Min(CSMF_j^{true}) \tag{B2}$$

$$CSMFAccuracy = 1 - \frac{\sum_{j=1}^{k} \left| CSMF_j^{true} - CSMF_j^{pred} \right|}{CSMFMaximumError} \tag{B3}$$

B.3.2 Partial Chance Corrected Concordance (PCCC)

PCCC measures the performance of models at the individual-level, comparing COD assignments between the physicians and models on a record by record basis, correcting for COD assignments made purely by chance [56]. PCCC is given by Equation B5, where k is the number of top COD assignments from the model to consider, N is number of causes, and C is fraction of records where the physician COD assignment is one of the top COD assignments from the model. For this study, k was set to 1, making C equivalent to the fraction of true positives TP or records, where the physician COD assignment is equal to the model COD assignment as shown in Equation B4. Higher PCCC values closer to 1 indicate that model COD assignments are similar to physician COD assignments, while values closer to 0 indicate that they are not similar to physicians.

$$C = \frac{TP}{Records} \tag{B4}$$

$$PCCC(k) = \frac{C - \frac{k}{N}}{1 - \frac{k}{N}} \tag{B5}$$

Appendix C Experiment on Repeated Runs of GPT-3.5

A short experiment was conducted to test the consistency of GPT-3.5 outputs repeated on the same record. 100 records, sampled randomly with approximately equal proportions across age groups, CODs, and survey rounds 1 and 2, were used to test repeated runs of GPT-3.5. Each record from the 100 records was rerun 10

462 times through GPT3.5, resulting in ten COD outputs per record. The ICD-10 codes were then converted to
463 CGHR-10 codes and tested for consistency, where completely inconsistent results had different ICD-10 or
464 CGHR-10 codes for each of the 10 reruns (1 times+), and completely consistent results had the same ICD-
465 10 or CGHR-10 code for all 10 reruns (10 times), on the same record.
466 The results are shown in Table C6. For all 100 records, GPT-3.5 assigns the same ICD-10 and CGHR-10 code
467 for the same record 5+ times out of 10. For 66 and 79 records, GPT-3.5 assigns the same ICD-10 and CGHR-
468 10 code respectively for each record. This number increases to 94 (from 66) and 96 (from 79) when
469 reducing the number of times out of 10 that GPT-3.5 assigns the same ICD-10 and CGHR-10 code
470 respectively. Thus, GPT-3.5 does not always produce the same outputs when repeated on the same record
471 (10 times out of 10), even when the temperature is set to 0, but does so for more than half the records. For
472 most records (more than 90%), GPT-3.5 will produce the same outputs for the same record 7+ times of 10.
473 **Table C6** Records with same GPT-3.5 outputs based on 10 repeated reruns of 100 records

Times with Same GPT-3.5 Outputs	ICD-10 Records	CGHR-10 Records
1 times+ (inconsistent)	100	100
2 times+	100	100
3 times+	100	100
4 times+	100	100
5 times+	100	100
6 times+	94	96
7 times+	92	94
8 times+	86	91
9 times+	79	86
10 times (consistent)	66	79

474
475

References

1. World Health Organization (2019) Non communicable diseases: Key Facts. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. Accessed 4 Jan 2024

2. Benziger CP, Roth GA, Moran AE (2016) The Global Burden of Disease Study and the Preventable Burden of NCD. Global Heart 11:393–397

3. Lawn JE, Kerber K, Enweronu-Laryea C, Cousens S (2010) 3.6 Million Neonatal Deaths—What Is Progressing and What Is Not? Seminars in Perinatology 34:371–386

4. Lassi ZS, Bhutta ZA (2015) Community-based intervention packages for reducing maternal and neonatal morbidity and mortality and improving neonatal outcomes. Cochrane Database of Systematic Reviews. https://doi.org/10.1002/14651858.CD007754.pub3

5. Liu NH, Daumit GL, Dua T, et al (2017) Excess mortality in persons with severe mental disorders: a multilevel intervention framework and priorities for clinical practice, policy and research agendas. World Psychiatry 16:30–40

6. Ewig S, Torres A (2011) Community-acquired pneumonia as an emergency: time for an aggressive intervention to lower mortality. European Respiratory Journal 38:253–260

7. World Health Organization (2021) SCORE for health data technical package: global report on health data systems and capacity, 2020.

8. de Savigny D, Riley I, Chandramohan D, et al (2017) Integrating community-based verbal autopsy into civil registration and vital statistics (CRVS): system-level considerations. Global Health Action 10:1272882

9. Thomas L-M, D'Ambruoso L, Balabanova D (2018) Verbal autopsy in health policy and systems: a literature review. BMJ Global Health 3:e000639

10. Rampatige R, Mikkelsen L, Hernandez B, Riley I, Lopez AD (2014) Systematic review of statistics on causes of deaths in hospitals: strengthening the evidence for policy-makers. Bull World Health Organ 92:807–816

11. Adair T (2021) Who dies where? Estimating the percentage of deaths that occur at home. BMJ Global Health 6:e006766

12. World Health Organization (2023) Verbal autopsy standards: 2022 WHO verbal autopsy instrument.

13. Chandramohan D, Fottrell E, Leitao J, et al (2021) Estimating causes of death where there is no medical certification: evolution and state of the art of verbal autopsy. Global Health Action 14:1982486

14. World Health Organization (2007) Verbal autopsy standards: ascertaining and attributing cause of death. World Health Organization

15. Gomes M, Begum R, Sati P, Dikshit R, Gupta PC, Kumar R, Sheth J, Habib A, Jha P (2017) Nationwide Mortality Studies To Quantify Causes Of Death: Relevant Lessons From India's Million Death Study. Health Affairs 36:1887–1895

16. Jha P, Gajalakshmi V, Gupta PC, Kumar R, Mony P, Dhingra N, Peto R, Collaborators R-CPS (2005) Prospective Study of One Million Deaths in India: Rationale, Design, and Validation Results. PLOS Medicine 3:e18

17. McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark SJ (2016) Probabilistic Cause-of-death Assignment using Verbal Autopsies. J Am Stat Assoc 111:1036–1049

18. Morris SK, Bassani DG, Kumar R, Awasthi S, Paul VK, Jha P (2010) Factors associated with physician agreement on verbal autopsy of over 27000 childhood deaths in India. PloS one 5:e9583

19. Soleman N, Chandramohan D, Shibuya K (2006) Verbal autopsy: current practices and challenges. Bulletin of the World Health Organization 84:239–245

20. Byass P, Hussain-Alkhateeb L, D'Ambruoso L, et al (2019) An integrated approach to processing WHO-2016 verbal autopsy data: the InterVA-5 model. BMC Medicine 17:102

21. Jha P, Kumar D, Dikshit R, et al (2019) Automated versus physician assignment of cause of death for verbal autopsies: randomized trial of 9374 deaths in 117 villages in India. BMC Medicine 17:116

22. Leitao J, Desai N, Aleksandrowicz L, et al (2014) 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 12:22

23. Desai N, Aleksandrowicz L, Miasnikof P, et al (2014) 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 12:20

24. Tunga M, Lungo J, Chambua J, Kateule R (2021) Verbal autopsy models in determining causes of death. Tropical Medicine & International Health 26:1560–1567

25. Oti SO, Kyobutungi C (2010) Verbal autopsy interpretation: a comparative analysis of the InterVA model versus physician review in determining causes of death in the Nairobi DSS. Popul Health Metrics 8:21

26. Jeblee S, Gomes M, Jha P, Rudzicz F, Hirst G (2019) Automatically determining cause of death from verbal autopsy narratives. BMC Med Inform Decis Mak 19:127

27. Blanco A, Pérez A, Casillas A, Cobos D (2021) Extracting Cause of Death From Verbal Autopsy With Deep Learning Interpretable Methods. IEEE Journal of Biomedical and Health Informatics 25:1315–1325

28. King C, Zamawe C, Banda M, et al (2016) The quality and diagnostic value of open narratives in verbal autopsy: a mixed-methods analysis of partnered interviews from Malawi. BMC Med Res Methodol 16:13

29. Chang Y, Wang X, Wang J, et al (2023) A Survey on Evaluation of Large Language Models. https://doi.org/10.48550/arXiv.2307.03109

30. Lund BD, Wang T (2023) Chatting about ChatGPT: how may AI and GPT impact academia and libraries? Library Hi Tech News 40:26–29

31. Svyatkovskiy A, Deng SK, Fu S, Sundaresan N (2020) 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. Association for Computing Machinery, New York, NY, USA, pp 1433–1443

32. Haupt CE, Marks M (2023) AI-Generated Medical Advice—GPT and Beyond. JAMA 329:1349–1350

33. Wu T, He S, Liu J, Sun S, Liu K, Han Q-L, Tang Y (2023) A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. IEEE/CAA JAS 10:1122–1136

34. OpenAI, Achiam J, Adler S, et al (2023) GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774

35. Njala University (2023) Healthy Sierra Leone. https://healsl.org/. Accessed 7 Jan 2024

36. Carshon-Marsh R, Aimone A, Ansumana R, et al (2022) Child, maternal, and adult mortality in Sierra Leone: nationally representative mortality survey 2018–20. The Lancet Global Health 10:e114–e123

37. World Health Organization (2011) ICD-10: International statistical classification of diseases and related health problems (10th revision).

38. Aleksandrowicz L, Malhotra V, Dikshit R, et al (2014) Performance criteria for verbal autopsy-based systems to estimate national causes of death: development and application to the Indian Million Death Study. BMC Medicine 12:21

39. Barnett ML, Boddupalli D, Nundy S, Bates DW (2019) Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. JAMA Network Open 2:e190096

40. Hsiao M, Morris SK, Bassani DG, Montgomery AL, Thakur JS, Jha P (2012) Factors Associated with Physician Agreement on Verbal Autopsy of over 11500 Injury Deaths in India. PLOS ONE 7:e30336

41. Murray CJ, Lozano R, Flaxman AD, et al (2014) Using verbal autopsy to measure causes of death: the comparative performance of existing methods. BMC Med 12:5

42. Benara SK, Sharma S, Juneja A, Nair S, Gulati BK, Singh KhJ, Singh L, Yadav VP, Rao C, Rao MVV (2023) Evaluation of methods for assigning causes of death from verbal autopsies in India. Front Big Data 6:1197471

43. Li ZR, Thomas J, Choi E, McCormick TH, Clark SJ (2023) The openVA Toolkit for Verbal Autopsies. The R Journal 1

44. Byass P, Chandramohan D, Clark SJ, et al (2012) Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool. Global Health Action 5:19281

45. BAYES (1958) An essay towards solving a problem in the doctrine of chances. Biometrika 45:296–315

46. Brooks S (1998) Markov chain Monte Carlo method and its application. J Royal Statistical Soc D 47:69–100

47. Chib S (2001) Markov chain Monte Carlo methods: computation and inference. Handbook of econometrics 5:3569–3649

48. Han C, Carlin BP (2001) Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review. Journal of the American Statistical Association 96:1122–1132

49. Brown TB, Mann B, Ryder N, et al (2020) Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

50. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is All you Need. Advances in Neural Information Processing Systems 30:

51. Ouyang L, Wu J, Jiang X, et al (2022) Training language models to follow instructions with human feedback. https://doi.org/10.48550/arXiv.2203.02155

52. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D (2017) Deep reinforcement learning from human preferences. Advances in neural information processing systems 30:

53. Stiennon N, Ouyang L, Wu J, Ziegler D, Lowe R, Voss C, Radford A, Amodei D, Christiano PF (2020) Learning to summarize with human feedback. Advances in Neural Information Processing Systems 33:3008–3021

54. Wirth C, Akrour R, Neumann G, Fürnkranz J (2017) A survey of preference-based reinforcement learning methods. J Mach Learn Res 18:4945–4990

55. World Health Organization (2016) Verbal autopsy standards: The 2016 WHO verbal autopsy instrument. https://www.who.int/publications/m/item/verbal-autopsy-standards-the-2016-who-verbal-autopsy-instrument. Accessed 17 Jan 2024

56. Murray CJ, Lozano R, Flaxman AD, Vahdatpour A, Lopez AD (2011) Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. Population Health Metrics 9:28

57. Setel PW, Whiting DR, Hemed Y, Chandramohan D, Wolfson LJ, Alberti KGMM, Lopez AD (2006) Validity of verbal autopsy procedures for determining cause of death in Tanzania. Tropical Medicine & International Health 11:681–696

58. Rasmussen LA, Cascio MA, Ferrand A, Shevell M, Racine E (2019) The complexity of physicians' understanding and management of prognostic uncertainty in neonatal hypoxic-ischemic encephalopathy. J Perinatol 39:278–285

59. Faison G, Chou F-S, Feudtner C, Janvier A (2023) When the Unknown Is Unknowable: Confronting Diagnostic Uncertainty. Pediatrics 152:e2023061193

60. Ansumana R, Mohamed V, Carshon-Marsh R, et al (2023) Report on Causes of Death in Sierra Leone 2018 – 2023.

61. Johnson D, Goodman R, Patrinely J, et al (2023) Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq rs.3.rs-2566942

62. Jang ME, Lukasiewicz T (2023) Consistency Analysis of ChatGPT.

63. Krishna S, Bhambra N, Bleakney R, Bhayana R, Atzen S (2024) Evaluation of Reliability, Repeatability, Robustness, and Confidence of GPT-3.5 and GPT-4 on a Radiology Board–style Examination. Radiology 311:e232715

64. Hoi SC, Sahoo D, Lu J, Zhao P (2021) Online learning: A comprehensive survey. Neurocomputing 459:249–289

65. Kavikondala A, Muppalla V, Prakasha DK, Acharya V (2019) Automated retraining of machine learning models. International Journal of Innovative Technology and Exploring Engineering 8:445–452

66. Wang S, Zhu Y, Liu H, Zheng Z, Chen C, Li J (2024) Knowledge Editing for Large Language Models: A Survey. ACM Comput Surv 57:59:1-59:37

67. Khowaja SA, Khuwaja P, Dev K, Wang W, Nkenyereye L (2024) ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) Evaluation: A Review. Cogn Comput. https://doi.org/10.1007/s12559-024-10285-1

68. Wu X, Duan R, Ni J (2024) Unveiling security, privacy, and ethical concerns of ChatGPT. Journal of Information and Intelligence 2:102–115

69. Intersoft Consulting (2018) General Data Protection Regulation (GDPR) – Legal Text. In: General Data Protection Regulation (GDPR). https://gdpr-info.eu/. Accessed 24 Jun 2025

70. Beck EJ, Gill ,Wayne, and De Lay PR (2016) Protecting the confidentiality and security of personal health information in low- and middle-income countries in the era of SDGs and Big Data. Global Health Action 9:32089

71. Kwarkye TG (2025) "We know what we are doing": the politics and trends in artificial intelligence policies in Africa. Canadian Journal of African Studies / Revue canadienne des études africaines 1–19

72. Das BC, Amini MH, Wu Y (2025) Security and Privacy Challenges of Large Language Models: A Survey. ACM Comput Surv 57:1–39

73. Wang F, Lin M, Ma Y, Liu H, He Q, Tang X, Tang J, Pei J, Wang S (2025) 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 V.2. ACM, Toronto ON Canada, pp 6173–6183

74. Corradini F, Leonesi M, Piangerelli M (2025) State of the Art and Future Directions of Small Language Models: A Systematic Review. Big Data and Cognitive Computing 9:189

75. Shawon MdTH, Ashrafi SAA, Azad AK, Firth SM, Chowdhury H, Mswia RG, Adair T, Riley I, Abouzahr C, Lopez AD (2021) Routine mortality surveillance to identify the cause of death pattern for out-of-hospital adult (aged 12+ years) deaths in Bangladesh: introduction of automated verbal autopsy. BMC Public Health 21:491

76. Maqungo M, Nannan N, Nojilana B, et al (2024) Can verbal autopsies be used on a national scale? Key findings and lessons from South Africa's national cause-of-death validation study. Global Health Action 17:2399413

77. Onyango D, Awuonda B (2024) Using verbal autopsy to enhance mortality surveillance. The Lancet Global Health 12:e1217–e1218

78. OpenAI (2024) Pricing. In: OpenAI. https://openai.com/api/pricing/. Accessed 4 Jul 2024

79. Liu X, Liu H, Yang G, Jiang Z, Cui S, Zhang Z, Wang H, Tao L, Sun Y, Song Z (2025) A generalist medical language model for disease diagnosis assistance. Nature medicine 31:932–942

80. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, Singhal K, Sharma Y, Azizi S, Kulkarni K (2025) Towards accurate differential diagnosis with large language models. Nature 1–7

81. World Health Organization (2022) ODK for verbal autopsy: A quick guide. https://www.who.int/publications/m/item/odk-for-verbal-autopsy--a-quick-guide. Accessed 24 Jan 2024

82. Nafundi (2023) ODK - Collect data anywhere.

83. DiPasquale A, Maire N, Bratschi M (2016) Release ODK 2016 WHO VA instrument 1.5.1 SwissTPH/WHO-VA.

84. Byass P (2020) InterVA-5.1 User Guide.

85. Thomas J, ekarpinskiMITRE, pkmitre, owentrigueros, Choi P, Chu Y pycrossva: Prepare data from WHO and PHRMC instruments for verbal autopsy algorithms.

86. OpenAI (2024) OpenAI Platform: API Reference (temperature parameter). https://platform.openai.com/docs/api-reference/completions/create#completions-create-temperature. Accessed 26 Jan 2024

87. Li ZR, McCormick T, Clark S (2022) InSilicoVA: Probabilistic Verbal Autopsy Coding with "InSilicoVA" Algorithm.

88. Thomas J, Li Z, Byass P, McCormick T, Boyas M, Clark S (2021) InterVA5: Replicate and Analyse "InterVA5."

89. Yendewa GA, Poveda E, Yendewa SA, Sahr F, Quiñones-Mateu ME, Salata RA (2018) HIV/AIDS in Sierra Leone: Characterizing the Hidden Epidemic. AIDS reviews 20:

90. Walker PG, White MT, Griffin JT, Reynolds A, Ferguson NM, Ghani AC (2015) Malaria morbidity and mortality in Ebola-affected countries caused by decreased health-care capacity, and the potential effect of mitigation strategies: a modelling analysis. The Lancet Infectious Diseases 15:825–832