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): richard.wen@utoronto.ca; Contributing authors: antenehta@gmail.com; andylee@cs.toronto.edu; 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) and InSilicoVA/InterVA-5 (0.44). GPT-4 achieved the highest PCCC for adult and neonatal deaths (0.64 and 0.58), with GPT-3.5 for child deaths (0.54). Across ages, model performance increased from 1 month to 14 years (∼0.10–0.75 PCCC) and declined from 15 to 69 years (∼0.70–0.35). GPT-4, 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 methodological details are provided in Appendix A.

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 2). 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

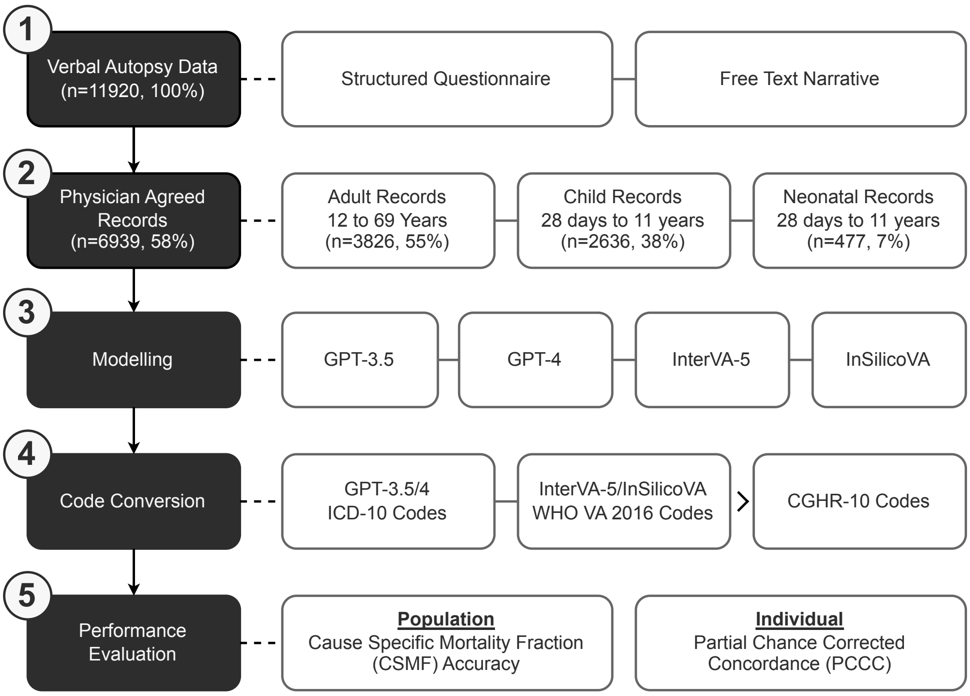

Fig. 1 Study methods.

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 Additional File 1), 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 A.1, with COD and age group distributions summarized in Tables A1 and A2.

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) based on transformer architectures [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 introduces 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 A.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 A.3.1), while Partial Chance-Corrected Concordance (PCCC) was used to assess individual-level agreement (see Appendix A.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 A.3.

3 Results

This section presents the performance of GPT-3.5, GPT-4, InterVA-5, and InSilicoVA in assigning CGHR-10 CODs, based on the methodology described in Section 2. GPT-4 achieved the highest overall individual-level concordance, with a PCCC of 0.61, followed by GPT-3.5 (0.56). GPT-4 also demonstrated the highest PCCC across most age groups and CODs within the adult (12–69 years), child (28 days–11 years), and neonatal (under 28 days) categories. In contrast, GPT-3.5, InterVA-5, and InSilicoVA showed higher PCCC values for a limited subset of age groups and CODs. Summary results are presented in Section 3.1, with stratified results by age group detailed in Sections 3.2, 3.3, and 3.4.

3.1 Overall Performance

Of all 6939 records, GPT-4 (0.61 PCCC) had the highest individual performance followed by GPT-3.5 (0.56 PCCC), InSilicoVA (0.44 PCCC), and InterVA-5 (0.44 PCCC) (Figure 2). GPT-3.5 and GPT-4 had improvements ranging from 0.14-0.18 PCCC over InSilicoVA and InterVA-5, while GPT-4 slightly improved over GPT-3.5 by 0.05 PCCC. Population level performances were similar for all models (0.74-0.79 CSMF). Figure 3 shows the PCCC performance across three age groups (adult, child, and neonate). GPT-4 had the best individual performance for adult and neonatal records (0.64 and 0.58 PCCC), while GPT-3.5 had the best performance for child records (0.54 PCCC) with GPT-4 performing slightly worse (0.51 PCCC). InSilicoVA and InterVA-5 performed the worse for adult and child records ($\leq$0.5 PCCC), while GPT-3.5 performed the worse for neonatal records (0.42 PCCC). Performance varied less for child deaths (0.13 range) than for adult and neonatal deaths (0.24 and 0.22 range). Across ages, all models followed a similar pattern in individual performance (Figure 4), where PCCC trended upwards for 1 month to 14 years ($\sim$0.1-0.75), and downwards for ages 15 to 69 years ($\sim$0.7-0.35). The highest and lowest performances were observed for ages 12-29 years ($\sim$0.4-0.7) and 1-11 months ($\sim$0.1-0.35) respectively.

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

Figure 5 presents model performance across 17 adult CODs, excluding suicide due to a low sample size (n=3, <1%). GPT-4 achieved the highest individual level performance for 10 of 17 CODs (0.35–0.99 PCCC), followed by GPT-3.5 for 5 CODs (0.43–0.94

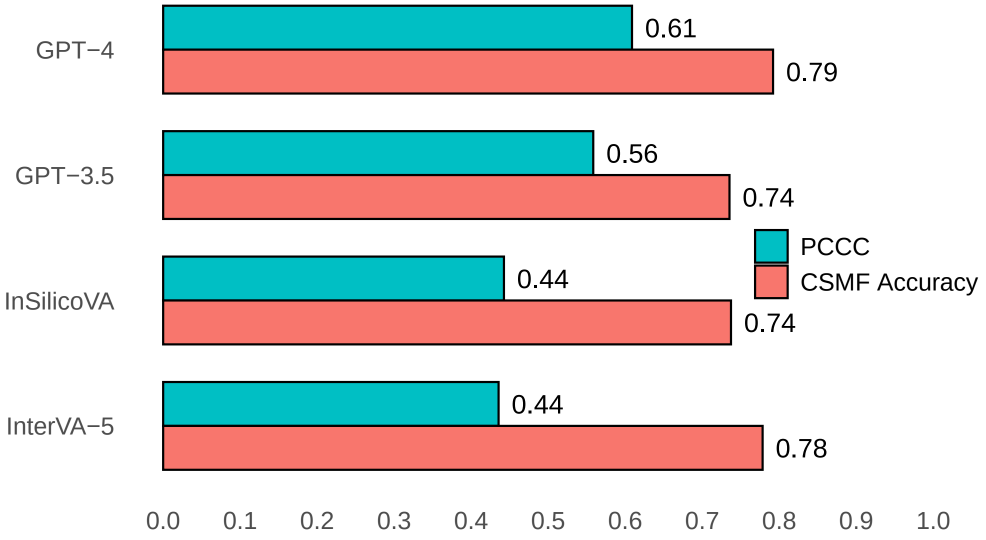

Fig. 2 Overall model performance.

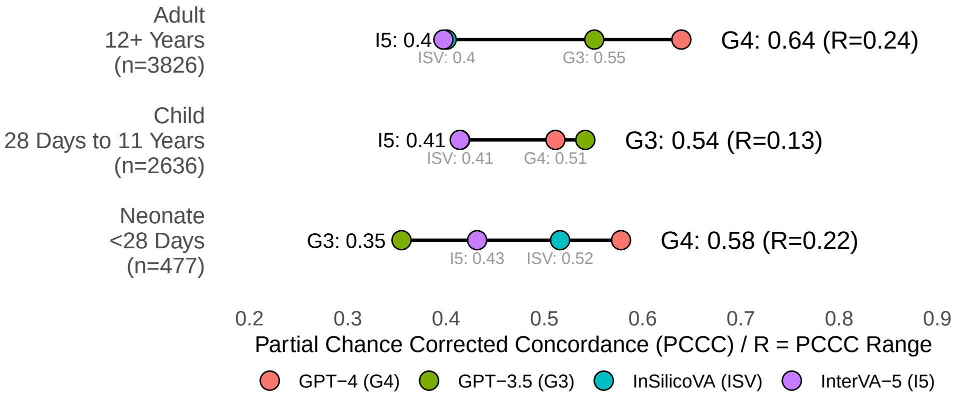

Fig. 3 Model performance by age group.

PCCC), and InSilicoVA for 2 CODs (0.71 and 0.84 PCCC). InterVA-5 showed the lowest performance for 8 CODs (0–0.79 PCCC), InSilicoVA for 6 CODs (0.01–0.41 PCCC), and GPT-3.5 for 2 CODs (0.38 and 0.53 PCCC). The greatest improvements of GPT-3.5/4 over InSilicoVA and InterVA-5 were observed in chronic respiratory

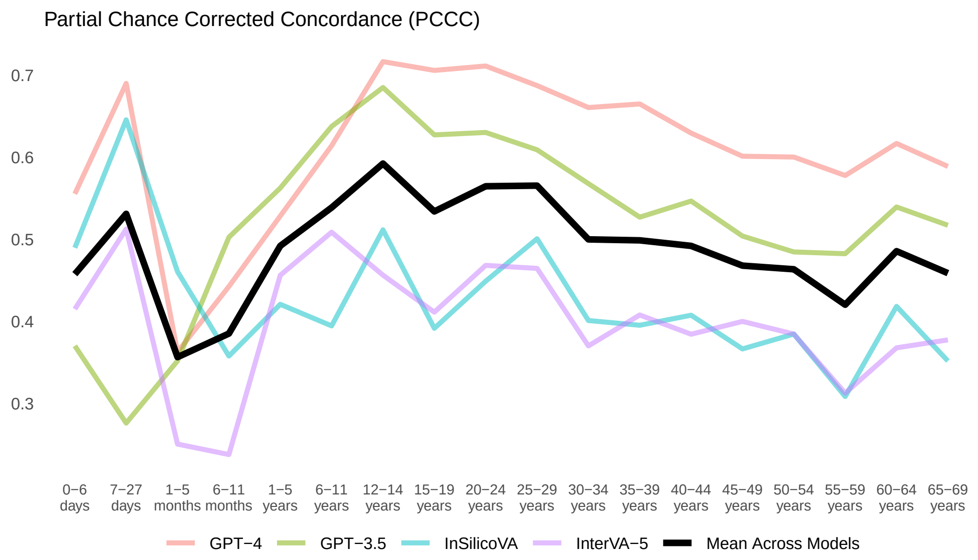

Fig. 4 Model performance by age range.

diseases (+0.74-0.94 PCCC), while the smallest improvements were for malaria (+0.09-0.17 PCCC). All models achieved PCCC values above 0.70 for maternal conditions (0.79–0.99), but remained below 0.50 for unspecified infections (0.35–0.49), malaria (0.26–0.43), and ill-defined CODs (0–0.35). GPT-4 showed performance improvements exceeding 0.20 PCCC over all other models for cancers (+0.25–0.36), stroke (+0.27–0.45), and diarrhoeal diseases (+0.37–0.51). GPT-3.5 demonstrated similar gains for liver and alcohol-related diseases (+0.27–0.52). Performance variability across models was most pronounced for chronic respiratory diseases (range: 0.94), while narrower differences were observed for maternal conditions (0.20), malaria (0.17), ischemic heart disease (0.15), and unspecified infections (0.14).

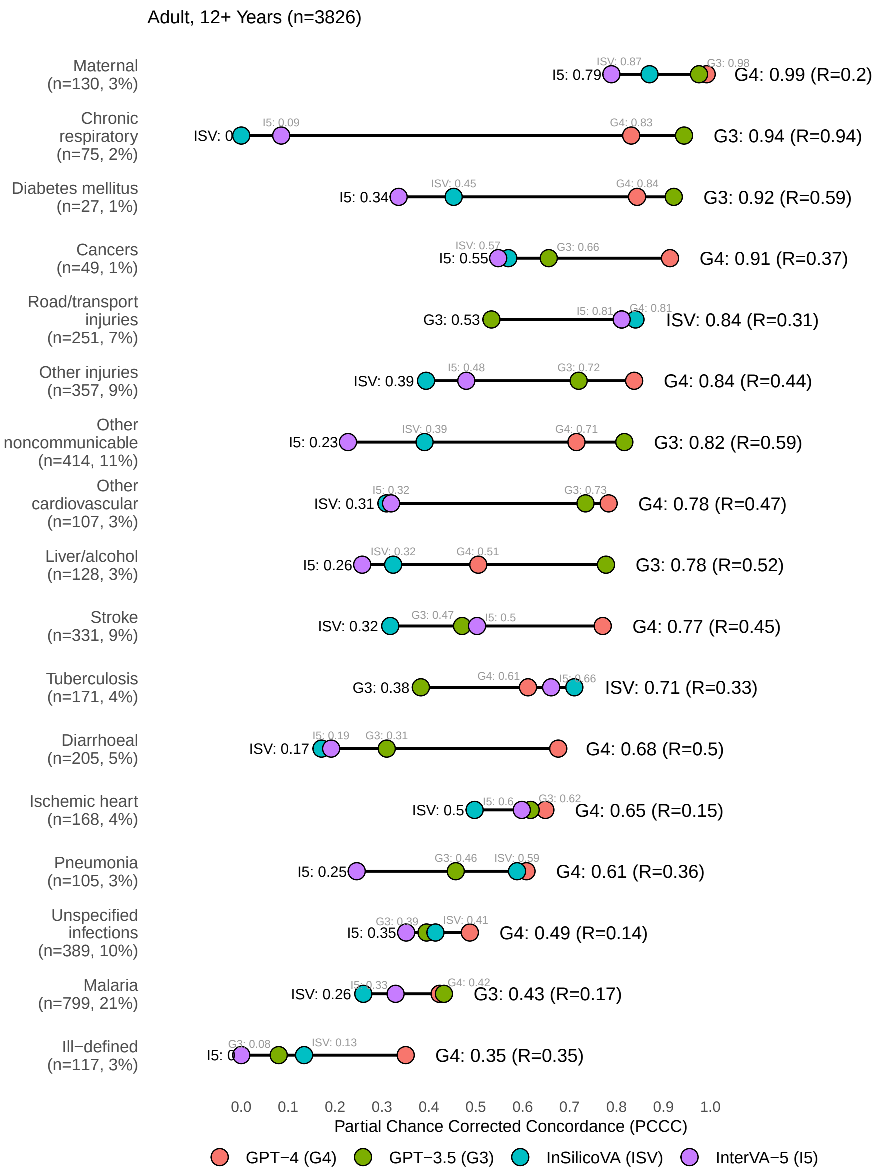

Fig. 5 Model performance for adult records by COD.

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, with PCCC values exceeding 0.70 (0.79–0.94), and showed 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 greater than 0.30 PCCC (+0.38–0.65), and also showed stronger performance for injuries, with gains of +0.11–0.15 compared to +0.01–0.04 for other models. Performance differences exceeding 0.60 PCCC were observed for epilepsy, leukaemia, other communicable diseases (range: 0.73), ill-defined causes (0.65), and nutritional deficiencies (0.61). In contrast, narrower differences (less than 0.30 PCCC) were seen for malaria (0.20), injuries, and other infections (0.15).

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 PCCC for 3 of the 5 CODs (0.39–0.71), while GPT-3.5 and InSilicoVA had the highest PCCC for 1 COD each (0.57 and 0.86). GPT-3.5 showed the lowest PCCC 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 PCCC), though only GPT-4 achieved a PCCC greater than 0 for prematurity-related deaths. InSilicoVA outperformed all other models for neonatal infections, with gains of +0.18–0.73 PCCC. Performance differences greater than 0.6 PCCC were observed for infections (range: 0.73) and prematurity and low birthweight (0.7). Stillbirth showed minimal variation across models (range: 0.09).

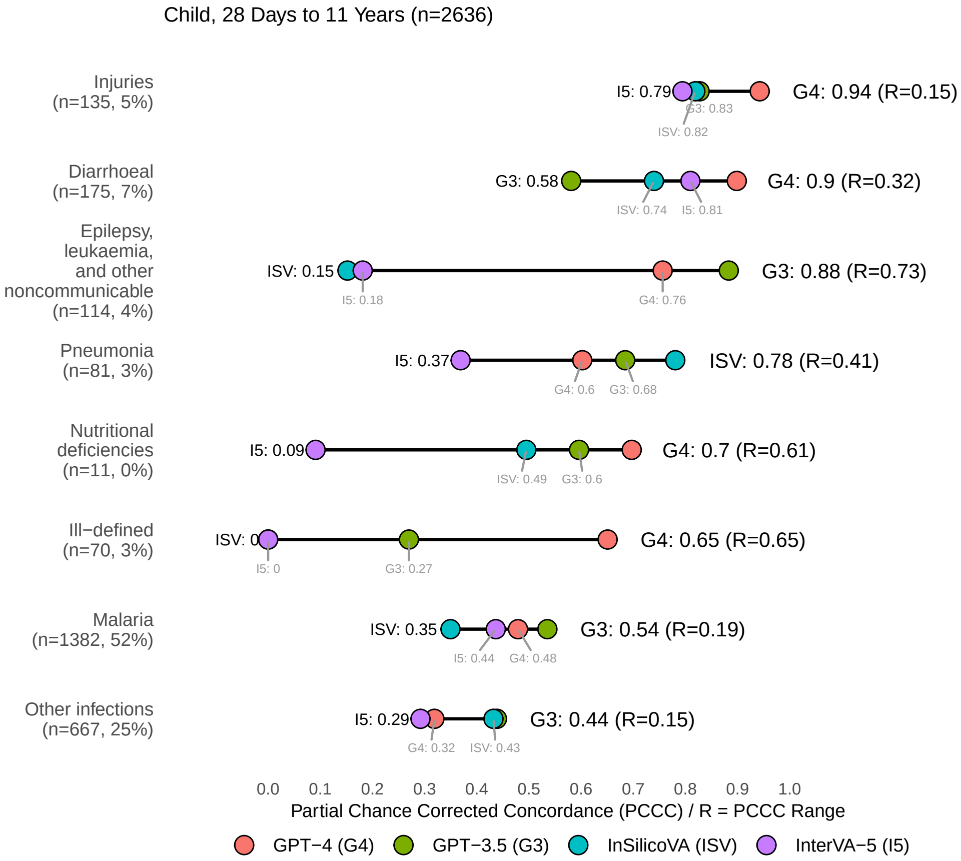

Fig. 6 Model performance for child records by COD.

4 Discussion

This section interprets and contextualizes the findings presented in Section 3. The comparative advantages and limitations of GPT-3.5, GPT-4, InterVA-5, and InSilicoVA for COD assignment are discussed in Sections 4.1 and 4.2, respectively. Study limitations are outlined in Section 4.3, and directions for future research are presented in Section 4.4.

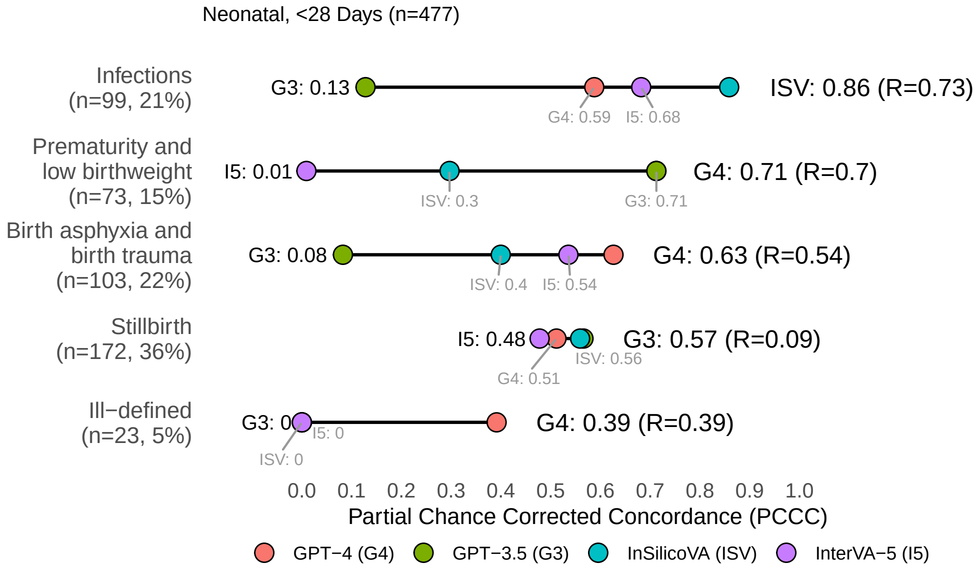

Fig. 7 Model performance for neonatal records by COD.

4.1 Advantages

This section outlines the strengths of the evaluated models in assigning CODs. Section 4.1.1 discusses model advantages across specific CODs and age groups. Section 4.1.3 highlights the potential for improving efficiency in physician-assisted COD assignment through computational support. Section 4.1.4 examines the benefits of leveraging natural language narratives in GPT models relative to traditional structured questionnaire data.

4.1.1 Cause-specific Models

At the population level, all models demonstrated comparable performance to physicians (0.74-0.79 CSMF), indicating their potential for estimating COD distributions in large populations. While individual-level performance was lower overall (0.44–0.61 PCCC), several models showed strong performance compared with physicians for specific CODs (up to 0.99 PCCC). GPT-3.5/4 consistently outperformed InSilicoVA and

InterVA-5 across most CODs, achieving the highest PCCC for 15 of 17 adult, 7 of 8 child, and 4 of 5 neonatal CODs. In contrast, InSilicoVA showed better performance for select CODs, including road and transport injuries (0.84 PCCC), tuberculosis (0.71), pneumonia (0.78), and neonatal infections (0.86). For CODs where high performance was observed, such as maternal conditions, chronic respiratory diseases, diabetes mellitus, and cancers for GPT-3.5/4 (0.91–0.99 PCCC), and road/transport injuries and neonatal infections for InSilicoVA (0.84 and 0.86 PCCC), the model outputs were more aligned with physician assignment. These findings support the potential utility of combining models based on their strengths for particular CODs. Evaluating performance at the COD level may allow for more targeted deployment of models, maximizing accuracy across disease categories. Table 1 illustrates how different models align with leading CODs identified in prior Sierra Leone studies [36, 58]. For example, we may deploy models to estimate asthma and chronic respiratory diseases using GPT-3 (0.94 PCCC), while using GPT-4 and InSilicoVA for diarrhoea and tuberculosis respectively (0.79 and 0.71 PCCC).

Table 1 Top ten leading causes of death for Sierra Leone in 2023 and most relevant models.

Top 10 Leading Cause of Death[1] (∼71% of ∼76K deaths)	Deaths (% of 76K)[2]	Best Model(s)	PCCC[3]
Malaria	16,075 (21%)	GPT-3.5/4	0.46 (n=2181)
Infections	11,777 (16%)	GPT-3.5/4/InSilicoVA	0.55 (n=1155)
Ischaemic heart and other vascular	5,747 (8%)	GPT-4	0.65 (n=168)
Diarrhoea	4,285 (6%)	GPT-4	0.79 (n=380)
Stroke	4,262 (6%)	GPT-4	0.77 (n=331)
Pneumonia	3,074 (4%)	GPT-4/InSilicoVA	0.7 (n=186)
Birth asphyxia and birth trauma	2,431 (3%)	GPT-4	0.63 (n=103)
Tuberculosis	2,399 (3%)	InSilicoVA	0.71 (n=171)
Low birth weight/preterm	1,570 (2%)	GPT-4	0.71 (n=103)
Asthma and chronic respiratory	1,551 (2%)	GPT-3	0.94 (n=75)

[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 [58]. Numbers are rounded.

[3] Adult, child, and neonate mean PCCC and summed n records if available.

4.1.2 Age-specific Performance Patterns

Across age groups, all models exhibited a consistent upward trend in performance from 6 months to 14 years, followed by a general decline from ages 15 to 69 years. GPT-3.5/4 outperformed InSilicoVA and InterVA-5 throughout this range, while performance patterns from birth to 5 months were more variable (see Figure 4). In adults, performance generally decreased with age, suggesting greater difficulty in assigning CODs among older adults, with a modest improvement observed after age 59. Among children and neonates, performance increased beyond 5 months, indicating greater model reliability as developmental age advanced. Although no model consistently achieved performances greater than 0.8 PCCC in any specific five-year age band, these age-related trends provide valuable insights. Specifically, they align with expectations from clinical literature, where physicians often face greater diagnostic uncertainty in neonatal cases [59, 60]. The observed patterns underscore the importance of considering developmental stage when interpreting model outputs and comparing them to physician-assigned CODs.

4.1.3 Scalability and Availability

The models evaluated in this study offer scalable and cost-effective support for physician-assigned CODs, particularly in resource-constrained settings. Similar to tools used in differential diagnosis, GPT and InSilicoVA models can provide alternative COD suggestions for physician review [39], potentially reducing the proportion of ill-defined causes and physician disagreement. At the time of analysis, running GPT-3.5 on 6,939 records cost approximately $1.60 USD (based on $0.50 per million tokens), while GPT-4 cost approximately $115 USD (at $30 per million tokens) [61]. InterVA-5 and InSilicoVA were freely available as open-source software. These costs compare favorably to physician review, which may exceed $3 USD per household in settings like India [15, 16], while the models can also process over 10,000 records within a single day. When physicians are unavailable, these models present a viable alternative for estimating population-level CODs. However, their application should be targeted to CODs where model performance is strong (see Table 1). Additionally, model outputs may be used to prioritize physician review, allocating less physician time to validating high-performing CODs (e.g. maternal conditions with 0.79–0.99 PCCC) and allocating more time to challenging cases (e.g. acute respiratory infections with 0.25–0.61 PCCC).

4.1.4 Natural Language Input and Output

None of the models required training data for COD assignment, enabling their use without domain-specific datasets or expertise. 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 are able to assign CODs using the ICD-10 standard, mirroring physician practice, and can potentially classify CODs in broader or alternative categories based on prompt design. In contrast, InterVA-5 and InSilicoVA rely exclusively on structured data from WHO VA 2016 questionnaires and assign CODs using WHO VA 2016 codes. This dependency necessitates ongoing maintenance and conversion between questionnaire versions (e.g., WHO VA 2012 to 2016) and coding systems (e.g., WHO VA 2016 to ICD-10), which reduces interoperability and comparability across 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 [26, 28]. For example, GPT-3.5/4 outperformed InterVA-5 and InSilicoVA by +0.35-0.65 PCCC on ill-defined CODs across age groups. They also demonstrated higher performance (+0.11-0.61 PCCC) on rarer CODs, such as nutritional deficiencies (n=11) and diabetes mellitus (n=27), which may be underrepresented in questionnaire data, but better contextualized through extensive knowledge embedded in GPT training corpora.

4.2 Disadvantages

This subsection addresses the caveats of GPT models in COD assignment. Section 4.2.1 examines challenges related to reproducibility of GPT outputs across repeated runs and their dependence on static training data. Section 4.2.2 explores the substantial computational resources required by GPT models and the associated concerns regarding data privacy and security.

4.2.1 Reproducibility and Timeliness

In this study, GPT models were run with the temperature parameter set to 0 to enhance reproducibility and consistency. However, a brief experiment (Appendix B) showed that GPT-3.5 assigned the same COD for the same record in just over 60% of repeated runs on a sample of 100 records. This variability indicates that GPT models do not consistently produce identical COD assignments for identical inputs, which raises concerns about reproducibility and reliability. For example, GPT models may correctly assign CODs by chance, but extensive testing with large numbers of reruns (e.g., 10,000) is cost-prohibitive, as rerunning increases costs substantially. By contrast, InterVA-5 and InSilicoVA are open-source and free, enabling unlimited reruns without additional expense. Moreover, these models provide COD assignments with probabilities for alternative causes, enhancing reproducibility and transparency despite lower overall performance. Another important limitation common to all models is their reliance on training data that reflect information only up to a fixed point in time. Consequently, they may not incorporate the most current data sources, such as recent scientific literature, social media, or emerging reports. This lag can limit their ability to detect new or emerging diseases (e.g., COVID-19) and shifts in COD distributions related to outbreaks or other public health changes unless regularly updated.

4.2.2 Infrastructure and Data Privacy

GPT-3.5/4 require substantial computational infrastructure for training and inference, making local deployment impractical due to cost and model ownership constraints. Consequently, sensitive data, such as identifiable personal information, must be transmitted to external servers, raising significant privacy concerns. Data submitted via prompts, which include narrative content used for COD assignment, may be collected by service providers (e.g., OpenAI) and potentially misused [62]. There is risk that sensitive information could be exposed or exploited through malicious actors or poorly controlled data handling [63, 64]. 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 [65–67]. In contrast, InterVA-5 and InSilicoVA can be run entirely on local systems, enabling data to remain under the control of the data owner. This approach reduces dependency on external services and better safeguards data privacy.

4.3 Limitations

This section outlines key limitations of the current study related to the use of GPT models. Section 4.3.1 discusses the use of physician assignment as the reference standard for comparing models. Section 4.3.2 addresses the need for further parameter tuning and the evaluation of model consistency and multi-COD assignments, while Section 4.3.3 notes the importance of more diverse datasets for model evaluation.

4.3.1 Physician Reference Standard

This study evaluated model performance using broad CGHR-10 categories rather than specific ICD-10 codes. In practice, physicians assign more detailed ICD-10 codes, but InterVA-5 and InSilicoVA generate only broader WHO VA codes and cannot assign ICD-10 codes directly, partly due to insufficient sample cases for many specific ICD-10 categories to support reliable modeling. For example, even broad CGHR-10 categories had fewer than 10 cases (e.g., congenital anomalies, suicide), and were excluded from evaluation. While GPT models assigned ICD-10 codes, lower performance can be expected, as even physicians show limited agreement on detailed ICD-10 coding, with only 6,939 (58%) of 11,920 records in agreement, necessitating the use of broader categories (e.g., CMEA-10 codes) to assess equivalence. Reliance on physician assignment as the reference standard may introduce bias, as physician interpretations may be shaped by local epidemiological knowledge, particularly for more complex cases or ambiguous narratives [22].

4.3.2 Model Tuning, Consistency, and Multiple Outputs

GPT-3.5/4 were used with default parameters except for temperature, which was set to 0 to enhance consistency. However, tuning temperature and other settings could potentially improve performance [68], but was not explored due to the high cost of repeated runs needed for sensitivity analyses, as noted in Section 4.2.1. Despite temperature control, GPT outputs may still vary, highlighting the need to assess reliability and consistency to avoid coincidental results [69–71]. Unlike GPT models, InterVA-5 and InSilicoVA provide multiple COD assignments with associated probabilities to measure reliability. In addition, while GPT can be prompted to generate multiple CODs, this study evaluated only the most probable assignment. Considering multiple COD outputs may better capture alternative diagnoses and align more closely with physician assessments [19].

4.3.3 Global Validity

While this study rigorously compares computer algorithms for COD assignment in Sierra Leone, the extent to which these findings are applied to other geographic or epidemiological contexts remains limited. Variations in local mortality profiles, linguistic expression, health system infrastructure, and culturally specific interpretations of illness shape the content and structure of VA narratives and questionnaires [72–74]. For example, the ways in which symptoms are described, terminology used, and aspects emphasized by respondents differ across languages and cultural settings. Moreover, Sierra Leone is predominantly driven by infectious diseases, such as malaria and respiratory infections, a pattern that contrasts with regions where non-communicable diseases typically constitute the leading CODs in North America and Europe, or where violence and road traffic injuries predominate in parts of Latin America and Asia [75–77]. 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 [78–80].

4.4 Opportunities

This section explores opportunities to enhance GPT models for assigning CODs. Section 4.4.1 highlights improvements through prompt engineering and analysis of misclassified cases. Section 4.4.2 discusses leveraging GPT to improve household survey data quality. Section 4.4.3 considers integrating GPT with InterVA-5 and InSilicoVA to support and enhance physician COD assignment within VA systems.

4.4.1 Prompt Engineering and Custom Models

Prompt engineering, the design of input prompts to guide GPT models toward improved outputs [81], offers a key opportunity to enhance COD assignment performance. An exploratory analysis in Appendix C of misclassified GPT-4 records for neonatal infections identified potential issues related to CGHR-10 code categorization, narrative information order, and COD assignment guidelines. Collaborating with domain experts (e.g., physicians, specialists) to review misclassified cases could inform prompt refinements that increase correct COD assignments or better align with broader COD categories. Furthermore, iterative prompt adjustments incorporating additional questionnaire data and physician manuals (e.g., via retrieval-augmented generation [82]) may improve model accuracy [83]. Sensitivity analyses can evaluate how prompt modifications affect performance and output consistency on a cause-specific basis. Additionally, GPT models can be customized to specific domains or contexts, adjusting objectives, behavior, data inputs, privacy considerations, and evaluation criteria to create specialized models optimized for particular CODs or settings [84].

4.4.2 Guided and Monitored Household Surveys

Verbal autopsies involve surveyors visiting households to collect information about the deceased from family, friends, or community members. While standardized questionnaires are used, important latent information within free-text narratives often goes uncaptured [26, 28]. Narrative quality depends heavily on the surveyor's social skills, cultural understanding, emotional capacity, and medical knowledge, all of which influence data completeness and potential bias [19, 85]. GPT models may support surveyors by suggesting improved or overlooked questions during interviews to elicit richer narratives. Moreover, as these models can assign CODs in real-time, they offer the opportunity to monitor data quality during collection. For example, by comparing estimated COD distributions with expected patterns for specific regions as a form of immediate quality control, where surveyors may be required to undergo review when estimated and expected COD distributions diverge significantly.

4.4.3 Computer Assisted Verbal Autopsy (CAVA)

This study establishes a basis for integrating GPT, InterVA-5, and InSilicoVA models into VA systems to support physicians in assigning CODs. In dual-coded VA systems (Section 2.1), two physicians independently assign CODs for each record and review each other's assignments (reconciliation), while a senior physician adjudicates if disagreements persist. As noted in Section 4.1.3, presenting alternative COD suggestions from GPT and InSilicoVA models may reduce physician disagreement and the frequency of ill-defined records, allowing physicians to focus on more complex cases. Model-generated COD suggestions can be offered to physicians after their initial assignment, enabling reconsideration or confirmation of CODs (step 2 and option 2b in Figure 8). Future work will evaluate the impact of these suggestions on improving VA data quality, including increasing physician agreement and reducing ill-defined deaths. GPT-4, InterVA-5, and InSilicoVA suggestions have been incorporated into the ongoing HEAL-SL study [35], aiming to improve physician agreement and lower ill-defined COD assignments.

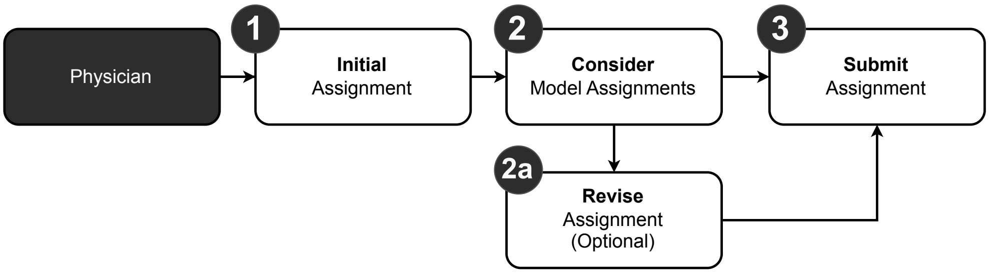

Fig. 8 Model suggestions integrated in the physician assignment process.

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 files were used to supplement this paper:

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

Acknowledgments. TBD.

Declarations

Funding

TBD.

Competing interests

Not applicable.

Ethics approval

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article (and its additional files), at https://openmortality.org (available upon request). Verbal Autopsy (VA) and narrative data by age group and survey rounds 1 and 2 available at https://openmortality.org/dataset/heal-sl. Cause of death code mappings to convert between ICD-10, WVA-2016, and CGHR-10 codes available at https://openmortality.org/dataset/icd.

Code availability

All code for this paper is available at https://github.com/cghr-toronto/healsl-gpt-paper.

Authors' contributions

PJ and PB are the study Principal Investigators. ATA and RK implemented the data collection procedures. RW, and TKSN 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.

Appendix A 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 A1 as a five-step process. Section A.1 provides details on the preprocessed data used for modelling. Section A.2 describes the data and parameter inputs and outputs for each model, while Section A.3 details the evaluation of model outputs at the individual and population level across different CODs, age groups, and ages.

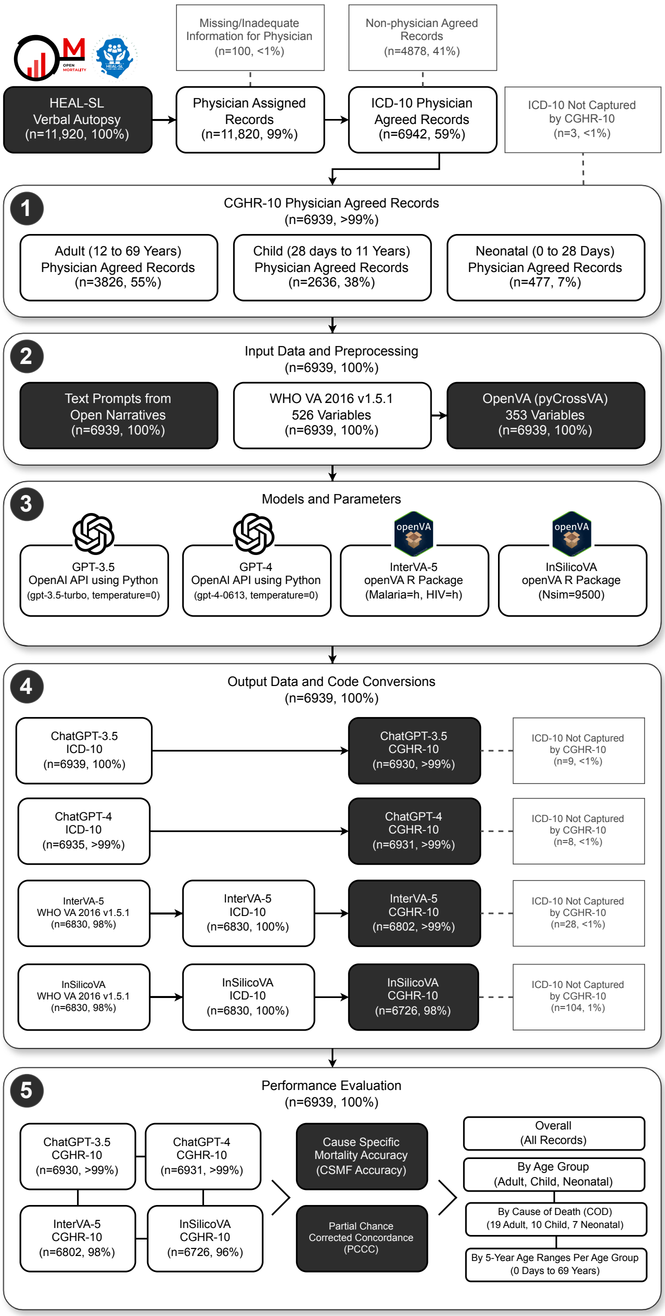
Missing/Inadequate Information for Physician (n=100, <1%)
Non-physician Agreed Records (n=4878, 41%)
HEAL-SL Verbal Autopsy (n=11,920, 100%)
Physician Assigned Records (n=11,820, 99%)
ICD-10 Physician Agreed Records (n=6942, 59%)
ICD-10 Not Captured by CGHR-10 (n=3, <1%)
1
CGHR-10 Physician Agreed Records (n=6939, >99%)
Adult (12 to 69 Years) Physician Agreed Records (n=3826, 55%)
Child (28 days to 11 Years) Physician Agreed Records (n=2636, 38%)
Neonatal (0 to 28 Days) Physician Agreed Records (n=477, 7%)
2
Input Data and Preprocessing (n=6939, 100%)
Text Prompts from Open Narratives (n=6939, 100%)
WHO VA 2016 v1.5.1 526 Variables (n=6939, 100%)
OpenVA (pyCrossVA) 353 Variables (n=6939, 100%)
3
Models and Parameters
GPT-3.5 OpenAI API using Python (gpt-3.5-turbo, temperature=0)
GPT-4 OpenAI API using Python (gpt-4-0613, temperature=0)
openVA
InterVA-5 openVA R Package (Malaria=h, HIV=h)
openVA
InSilicoVA openVA R Package (Nsim=9500)
4
Output Data and Code Conversions (n=6939, 100%)
ChatGPT-3.5 ICD-10 (n=6939, 100%)
ChatGPT-3.5 CGHR-10 (n=6930, >99%)
ICD-10 Not Captured by CGHR-10 (n=9, <1%)
ChatGPT-4 ICD-10 (n=6935, >99%)
ChatGPT-4 CGHR-10 (n=6931, >99%)
ICD-10 Not Captured by CGHR-10 (n=8, <1%)
InterVA-5 WHO VA 2016 v1.5.1 (n=6830, 98%)
InterVA-5 ICD-10 (n=6830, 100%)
InterVA-5 CGHR-10 (n=6802, >99%)
ICD-10 Not Captured by CGHR-10 (n=28, <1%)
InSilicoVA WHO VA 2016 v1.5.1 (n=6830, 98%)
InSilicoVA ICD-10 (n=6830, 100%)
InSilicoVA CGHR-10 (n=6726, 98%)
ICD-10 Not Captured by CGHR-10 (n=104, 1%)
5
Performance Evaluation (n=6939, 100%)
ChatGPT-3.5 CGHR-10 (n=6930, >99%)
ChatGPT-4 CGHR-10 (n=6931, >99%)
InterVA-5 CGHR-10 (n=6802, 98%)
InSilicoVA CGHR-10 (n=6726, 96%)
Cause Specific Mortality Accuracy (CSMF Accuracy)
Partial Chance Corrected Concordance (PCCC)
Overall (All Records)
By Age Group (Adult, Child, Neonatal)
By Cause of Death (COD) (19 Adult, 10 Child, 7 Neonatal)
By 5-Year Age Ranges Per Age Group (0 Days to 69 Years)

Fig. A1 Detailed study methods.

A.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 2). The 6942 records were converted into CGHR-10 codes (see Additional File 1) 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 A2. 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 A1. 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 noncommunicable diseases (61%), other injuries (77%), road and transport injuries (71%), and tuberculosis (68%). Most records were coded by physicians as malaria for adults (20%) and children (52%), and stillbirth (36%) and neonatal infections (21%)

for neonates. Suicide, congenital anomalies, nutritional deficiencies, and other had low sample sizes for each age group (<1% of total records for each age group). Table A2 shows the distribution of the study data by age. Across ages, there were more male records for 50-59 years (60-62%), while all other records had between 49-59% female and male records. Most records were in the 65-69 years age range for adults (15%), 1-5 years for children (62%), and 0-6 days for neonates (83%).

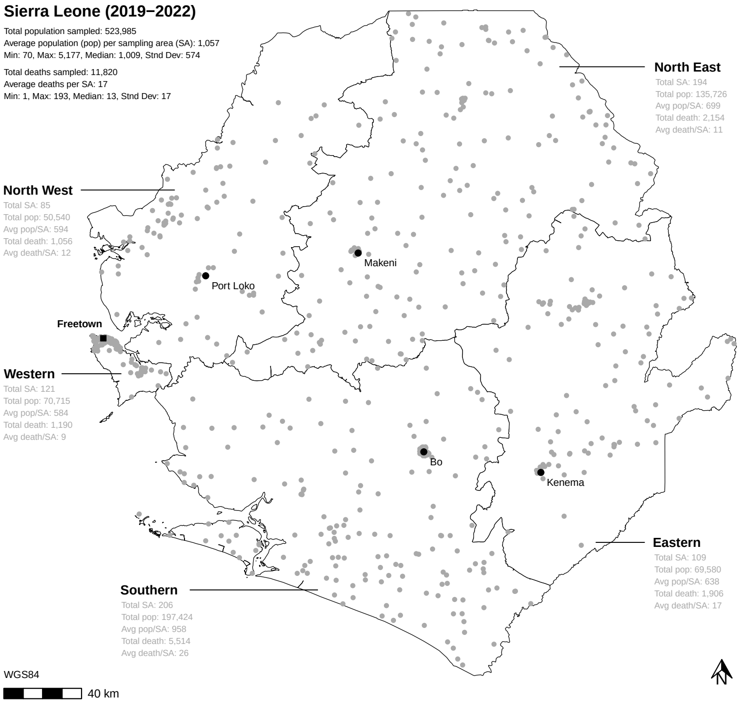

Fig. A2 Study data sampling areas.

A.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

Table A1 Study data by cause of death.

Age Group	CGHR-10 Cause of Death (COD)	Female	Male	Total
Adult, 18 CODs (n=3826, 55.1%) Adult Female (n=1681, 43.9%) Adult Male (n=2145, 56.1%)	Acute Respiratory Infections	48 (45.7%)	57 (54.3%)	105 (2.7%)
	Cancers	32 (65.3%)	17 (34.7%)	49 (1.3%)
	Chronic Respiratory Diseases	29 (38.7%)	46 (61.3%)	75 (2%)
	Diabetes Mellitus	14 (51.9%)	13 (48.1%)	27 (0.7%)
	Diarrhoeal Diseases	102 (49.8%)	103 (50.2%)	205 (5.4%)
	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%)
	Unspecified Infections	175 (45%)	214 (55%)	389 (10.2%)
Child, 9 CODs (n=2636, 38%) Child Female (n=1290, 48.9%) Child Male (n=1346, 51.1%)	Congenital Anomalies	1 (100%)	N/A	1 (0%)
	Diarrhoeal Diseases	79 (45.1%)	96 (54.9%)	175 (6.6%)
	Epilepsy, Leukaemia, And Other Noncommunicable Diseases	61 (53.5%)	53 (46.5%)	114 (4.3%)
	Ill-Defined	34 (48.6%)	36 (51.4%)	70 (2.7%)
	Injuries	51 (37.8%)	84 (62.2%)	135 (5.1%)
	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 (n=477, 6.9%) Neonate Female (n=227, 47.6%) Neonate Male (n=250, 52.4%)	Birth Asphyxia And Birth Trauma	38 (36.9%)	65 (63.1%)	103 (21.6%)
	Congenital Anomalies	2 (100%)	N/A	2 (0.4%)
	Ill-Defined	11 (47.8%)	12 (52.2%)	23 (4.8%)
	Neonatal Infections	49 (49.5%)	50 (50.5%)	99 (20.8%)
	Other	2 (40%)	3 (60%)	5 (1%)
	Prematurity And Low Birthweight	39 (53.4%)	34 (46.6%)	73 (15.3%)
	Stillbirth	86 (50%)	86 (50%)	172 (36.1%)

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 A.2.1 describes the input data and parameters for each model, while Section A.2.3 details the outputs from running each model.

Table A2 Study data by age range.

Age Group	Age Range	Female	Male	Total
Adult (n=3826, 55.1%) Adult Female (n=1681, 43.9%) Adult Male (n=2145, 56.1%)	12-14 Years	51 (37.8%)	84 (62.2%)	135 (3.5%)
	15-19 Years	115 (42.8%)	154 (57.2%)	269 (7%)
	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%) Child Female (n=1290, 48.9%) Child Male (n=1346, 51.1%)	1-5 Months	146 (47.4%)	162 (52.6%)	308 (11.7%)
	6-11 Months	160 (50.8%)	155 (49.2%)	315 (11.9%)
	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%) Neonate Female (n=227, 47.6%) Neonate Male (n=250, 52.4%)	0-6 Days	184 (46.6%)	211 (53.4%)	395 (82.8%)
	7-27 Days	43 (52.4%)	39 (47.6%)	82 (17.2%)

A.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 for each record:

```
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 [86, 87] consisting of 526 variables [88], followed by further conversion into OpenVA format [43] consisting of 353 variables [89] using the `pyCrossVA` version 0.97 Python package [90]. The 6939 records were all converted into OpenVA formatted records for InterVA-5 and InSilicoVA.

A.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 [68]. 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 [91] for InSilicoVA was set to 9500, while the `HIV` (level of prevalence of human immunodeficiency virus) and `Malaria` (level of prevalence of Malaria) parameters [92] for InterVA-5 were both set to `'h'` (high) reflecting HIV and Malaria disease assumptions in Sierra Leone [93, 94]. 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 for all iterations.

A.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.

A.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 A.3.1 describes CSMF accuracy in detail for evaluating models on the population level, Section A.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 to be 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.

A.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 A1. Then, the $CSMFMaximumError$, representing the worst possible model, is calculated using Equation A2. Finally, the CSMF accuracy is given by Equation A3, 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 at all.

$$CSMF_j = Records_j/Records \tag{A1}$$

$$CSMFMaximumError = 2(1 - Min(CSMF_j^{true}) \tag{A2}$$

$$CSMFAccuracy = 1 - \frac{\sum_{j=1}^{k} |CSMF_j^{true} - CSMF_j^{pred}|}{CSMFMaximumError} \tag{A3}$$

A.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 A5, 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 A4. Higher PCCC values closer to 1 indicate that model COD assignments are similar to physician COD assignments, while values closer to 0 indicate that model COD assignments are not similar to physicians.

$$C = \frac{TP}{Records} \tag{A4}$$

$$PCCC(k) = \frac{C - \frac{k}{N}}{1 - \frac{k}{N}} \tag{A5}$$

Appendix B 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 times through GPT-3.5, resulting in ten COD outputs per record. The ICD-10 codes were then converted to CGHR-10 codes and tested for consistency, where completely inconsistent results had different ICD-10 or CGHR-10 codes for each of the 10 reruns (1 times+), and completely consistent results had the same ICD-10 or CGHR-10 code for all 10 reruns (10 times), on the same record.

The results are shown in Table B3. For all 100 records, GPT-3.5 assigns the same ICD-10 and CGHR-10 code for the same record 5 times or more out of 10. For 66 and 79 records, GPT-3.5 assigns the same ICD-10 and CGHR-10 code respectively for each record. This number increases to 94 (from 66) and 96 (from 79) when reducing the number of times out of 10 that GPT-3.5 assigns the same ICD-10 and CGHR-10 code respectively. Thus, GPT-3.5 does not always produce the same outputs when repeated on the same record (10 times out of 10), even when the temperature is set to 0, but does so for more than half the records. For most records (more than 90%), GPT-3.5 will produce the same outputs for the same record 7 times or more out of 10.

Table B3 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

Appendix C Exploration of Neonatal Infections

An exploration of neonatal infections (n=99, 21% of 477 records) was done to understand the low performance of GPT models (0.23 PCCC) for neonatal infections, and high performance of InSilicoVA (0.87 PCCC). In Table C4, about half the records were assigned correctly, and a majority (n=33, 33%) of the other records were misclassified as other, while prematurity and low birthweight, birth asphyxia & birth trauma, and ill-defined make up the rest. On closer inspection of the 49 records with misclassified assignments, the ICD-10 code R50 was assigned in 20 records. R50 falls under unspecified infections in the adult CGHR-10 category, but in the other category for neonates. B50 was assigned in 4 records, falling under malaria, but a similar B54 falls under neonatal infections. P81 was assigned in 3 records, referring to fever of unknown origin, which falls under other, and P07 was assigned in 7 records, falling under prematurity and low birthweight.

In most misclassified records, there is mention of infections, but the misclassifications occur due to the finer details of the ICD-10 code classifications, the categorization decisions of the CGHR-10 codes, and missing information from the questionnaire. For R50 misclassifications, GPT may have confused descriptions across adult and neonatal age groups. Using the same definition of R50, but in the context of neonates, may result in codes closer to neonatal infections (e.g. B54). For B50 misclassifications, the similar B54 was categorized in CGHR-10 as neonatal infections, but B50 was categorized as other. P81 refers to fever of unknown origin, which may be difficult to differentiate between infection and other causes without information from the questionnaire. P07 refers to prematurity and low birthweight, where GPT initially assigned P07 as the age of the neonate was mentioned first, but later mentions infections as an alternative following the order of information in the narratives. Thus, it may be possible to improve the performance GPT models using better prompts based on the context of VA manuals and CGHR-10 codes, and by also including questionnaire information in the prompts.

Table C4 GPT-4 CGHR-10 COD assignment for physician coded neonatal infections records.

GPT-4 Assigned Cause of Death (CGHR-10)	Records
Neonatal infections	50 (51%)
Other	33 (33%)
Prematurity and low birthweight	9 (9%)
Birth asphyxia & birth trauma	5 (6%)
Ill-defined	2 (2%)
Total	99 (100%)

References

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

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

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

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

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

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

[7] World Health Organization. SCORE for Health Data Technical Package: Global Report on Health Data Systems and Capacity, 2020; 2021.

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

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

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

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

[12] World Health Organization. Verbal Autopsy Standards: 2022 WHO Verbal Autopsy Instrument; 2023.

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

[14] World Health Organization. Verbal Autopsy Standards: Ascertaining and Attributing Cause of Death. World Health Organization; 2007.

[15] Gomes M, Begum R, Sati P, Dikshit R, Gupta PC, Kumar R, et al. Nationwide Mortality Studies To Quantify Causes Of Death: Relevant Lessons From India's

Million Death Study. Health Affairs. 2017 Nov;36(11):1887–1895. https://doi.org/10.1377/hlthaff.2017.0635.

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

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

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

[19] Soleman N, Chandramohan D, Shibuya K. Verbal Autopsy: Current Practices and Challenges. Bulletin of the World Health Organization. 2006;84(3):239–245.

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

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

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

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

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

[25] Oti SO, Kyobutungi C. Verbal Autopsy Interpretation: A Comparative Analysis of the InterVA Model versus Physician Review in Determining Causes of Death in the Nairobi DSS. Population Health Metrics. 2010 Jun;8(1):21. https://doi.org/10.1186/1478-7954-8-21.

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

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

[28] King C, Zamawe C, Banda M, Bar-Zeev N, Beard J, Bird J, et al. The Quality and Diagnostic Value of Open Narratives in Verbal Autopsy: A Mixed-Methods Analysis of Partnered Interviews from Malawi. BMC Medical Research Methodology. 2016 Feb;16(1):13. https://doi.org/10.1186/s12874-016-0115-5.

[29] Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al.: A Survey on Evaluation of Large Language Models. arXiv.

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

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

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

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

[34] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al.: GPT-4 Technical Report. arXiv.

[35] Njala University.: Healthy Sierra Leone. https://healsl.org/.

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

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

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

[39] Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. JAMA Network Open. 2019 Mar;2(3):e190096. https://doi.org/10.1001/jamanetworkopen.2019.0096.

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

[41] Murray CJ, Lozano R, Flaxman AD, Serina P, Phillips D, Stewart A, et al. Using Verbal Autopsy to Measure Causes of Death: The Comparative Performance of Existing Methods. BMC Medicine. 2014 Jan;12(1):5. https://doi.org/10.1186/1741-7015-12-5.

[42] Benara SK, Sharma S, Juneja A, Nair S, Gulati BK, Singh KJ, et al. Evaluation of Methods for Assigning Causes of Death from Verbal Autopsies in India. Frontiers in Big Data. 2023 Aug;6:1197471. https://doi.org/10.3389/fdata.2023.1197471.

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

[44] Byass P, Chandramohan D, Clark SJ, D'Ambruoso L, Fottrell E, Graham WJ, et al. Strengthening Standardised Interpretation of Verbal Autopsy Data: The New InterVA-4 Tool. Global Health Action. 2012 Dec;5(1):19281. https://doi.org/10.3402/gha.v5i0.19281.

[45] BAYES. An Essay towards Solving a Problem in the Doctrine of Chances. Biometrika. 1958;45(3-4):296–315.

[46] Brooks S. Markov Chain Monte Carlo Method and Its Application. Journal of the Royal Statistical Society: Series D (The Statistician). 1998 Mar;47(1):69–100. https://doi.org/10.1111/1467-9884.00117.

[47] Chib S. Markov Chain Monte Carlo Methods: Computation and Inference. Handbook of econometrics. 2001;5:3569–3649.

[48] Han C, Carlin BP. Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review. Journal of the American Statistical Association. 2001 Sep;96(455):1122–1132. https://doi.org/10.1198/016214501753208780.

[49] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al.: Language Models Are Few-Shot Learners. arXiv.

[50] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. In: Advances in Neural Information Processing Systems.

vol. 30. Curran Associates, Inc.; 2017. .

[51] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al.: Training Language Models to Follow Instructions with Human Feedback. arXiv.

[52] Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep Reinforcement Learning from Human Preferences. Advances in neural information processing systems. 2017;30.

[53] Stiennon N, Ouyang L, Wu J, Ziegler D, Lowe R, Voss C, et al. Learning to Summarize with Human Feedback. Advances in Neural Information Processing Systems. 2020;33:3008–3021.

[54] Wirth C, Akrour R, Neumann G, Fürnkranz J. A Survey of Preference-Based Reinforcement Learning Methods. The Journal of Machine Learning Research. 2017 Jan;18(1):4945–4990.

[55] World Health Organization.: 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.

[56] 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;9(1):28. https://doi.org/10.1186/1478-7954-9-28.

[57] Setel PW, Whiting DR, Hemed Y, Chandramohan D, Wolfson LJ, Alberti KGMM, et al. Validity of Verbal Autopsy Procedures for Determining Cause of Death in Tanzania. Tropical Medicine & International Health. 2006;11(5):681–696. https://doi.org/10.1111/j.1365-3156.2006.01603.x.

[58] Ansumana R, Mohamed V, Carshon-Marsh R, Jambai A, Smart F, Sartie K, et al. Report on Causes of Death in Sierra Leone 2018 – 2023; 2023.

[59] Rasmussen LA, Cascio MA, Ferrand A, Shevell M, Racine E. The Complexity of Physicians' Understanding and Management of Prognostic Uncertainty in Neonatal Hypoxic-Ischemic Encephalopathy. Journal of Perinatology. 2019 Feb;39(2):278–285. https://doi.org/10.1038/s41372-018-0296-3.

[60] Faison G, Chou FS, Feudtner C, Janvier A. When the Unknown Is Unknowable: Confronting Diagnostic Uncertainty. Pediatrics. 2023 Sep;152(4):e2023061193. https://doi.org/10.1542/peds.2023-061193.

[61] OpenAI.: Pricing. https://openai.com/api/pricing/.

[62] Tao G, Cheng S, Zhang Z, Zhu J, Shen G, Zhang X.: Opening A Pandora's Box: Things You Should Know in the Era of Custom GPTs. arXiv.

[63] Khowaja SA, Khuwaja P, Dev K, Wang W, Nkenyereye L. ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital Divide, and Ethics) Evaluation: A Review. Cognitive Computation. 2024 May;https://doi.org/10.1007/s12559-024-10285-1.

[64] Wu X, Duan R, Ni J. Unveiling Security, Privacy, and Ethical Concerns of ChatGPT. Journal of Information and Intelligence. 2024;2(2):102–115.

[65] Intersoft Consulting.: General Data Protection Regulation (GDPR) – Legal Text. https://gdpr-info.eu/.

[66] Beck EJ, G Wayne, and De Lay PR. 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. 2016 Dec;9(1):32089. https://doi.

org/10.3402/gha.v9.32089.

[67] Kwarkye TG. “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. 2025;p. 1–19. https://doi.org/10.1080/00083968.2025.2456619.

[68] OpenAI.: OpenAI Platform: API Reference (Temperature Parameter). https://platform.openai.com/docs/api-reference/completions/create#completions-create-temperature.

[69] 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. Research Square. 2023 Feb;p. rs.3.rs–2566942. https://doi.org/10.21203/rs.3.rs-2566942/v1.

[70] Jang ME, Lukasiewicz T.: Consistency Analysis of ChatGPT. arXiv.

[71] 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. Radiology. 2024 May;311(2):e232715. https://doi.org/10.1148/radiol.232715.

[72] 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):1569–1577.

[73] Fottrell E, Byass P. Verbal Autopsy: Methods in Transition. Epidemiologic reviews. 2010;32(1):38–55.

[74] Byass P. The Imperfect World of Global Health Estimates. PLoS Medicine. 2010 Nov;7(11):e1001006. https://doi.org/10.1371/journal.pmed.1001006.

[75] Murray CJ, Lopez AD. Mortality by Cause for Eight Regions of the World: Global Burden of Disease Study. The lancet. 1997;349(9061):1269–1276.

[76] Mathers CD, Boerma T, Ma Fat D. Global and Regional Causes of Death. British medical bulletin. 2009;92(1):7–32.

[77] Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, et al. Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019. The lancet. 2020;396(10258):1204–1222.

[78] Shawon MTH, Ashrafi SAA, Azad AK, Firth SM, Chowdhury H, Mswia RG, et al. 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. 2021 Dec;21(1):491. https://doi.org/10.1186/s12889-021-10468-7.

[79] Maqungo M, Nannan N, Nojilana B, Nichols E, Morof D, Cheyip M, et al. 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. 2024 Dec;17(1):2399413. https://doi.org/10.1080/16549716.2024.2399413.

[80] Onyango D, Awuonda B. Using Verbal Autopsy to Enhance Mortality Surveillance. The Lancet Global Health. 2024;12(8):e1217–e1218.

[81] Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, et al.: Prompt Engineering for Healthcare: Methodologies and Applications. arXiv.

[82] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20. Red Hook, NY, USA: Curran Associates Inc.; 2020. p. 9459–9474.

[83] Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. Journal of medical Internet research. 2023;25:e50638.

[84] Almasre M. Development and Evaluation of a Custom GPT for the Assessment of Students' Designs in a Typography Course. Education Sciences. 2024 Feb;14(2):148. https://doi.org/10.3390/educsci14020148.

[85] Loh P, Fottrell E, Beard J, Bar-Zeev N, Phiri T, Banda M, et al. Added Value of an Open Narrative in Verbal Autopsies: A Mixed-Methods Evaluation from Malawi. BMJ Paediatrics Open. 2021 Feb;5(1):e000961. https://doi.org/10.1136/bmjpo-2020-000961.

[86] World Health Organization.: ODK for Verbal Autopsy: A Quick Guide. https://www.who.int/publications/m/item/odk-for-verbal-autopsy–a-quick-guide.

[87] Nafundi.: ODK - Collect Data Anywhere.

[88] DiPasquale A, Maire N, Bratschi M.: Release ODK 2016 WHO VA Instrument 1.5.1 SwissTPH/WHO-VA. Swiss Tropical and Public Health Institute.

[89] Byass P.: InterVA-5.1 User Guide.

[90] Thomas J, ekarpinskiMITRE, pkmitre, owentrigueros, Choi P, Chu Y.: Pycrossva: Prepare Data from WHO and PHRMC Instruments for Verbal Autopsy Algorithms.

[91] Li ZR, McCormick T, Clark S.: InSilicoVA: Probabilistic Verbal Autopsy Coding with 'InSilicoVA' Algorithm.

[92] Thomas J, Li Z, Byass P, McCormick T, Boyas M, Clark S.: InterVA5: Replicate and Analyse 'InterVA5'.

[93] Yendewa GA, Poveda E, Yendewa SA, Sahr F, Quiñones-Mateu ME, Salata RA. HIV/AIDS in Sierra Leone: Characterizing the Hidden Epidemic. AIDS reviews. 2018;20(2).

[94] Walker PG, White MT, Griffin JT, Reynolds A, Ferguson NM, Ghani AC. 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. 2015;15(7):825–832.