Comparing Generative Pre-trained Transformer Models to Physicians for Assigning Causes of Death from Verbal Autopsy: A Case Study of 6939 Deaths in Sierra Leone from 2019-2022

Richard Wen[1*], Anteneh Tesfaye Assalif[1,2], Andy Sze-Heng Lee[1], Rajeev Kamadod[1], Cheryl Chin[1], Asha Behdinan[1], Leslie Newcombe[1], Areeba Zubair[1], Thomas Kai Sze Ng[1], Paijani Sheth[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; cheryl.chin@unityhealth.to; asha.behdinan@mail.utoronto.ca; leslie.newcombe@unityhealth.to; areeba.zubair@mail.utoronto.ca; kaisze.ng@unityhealth.to; paijani.sheth@mail.utoronto.ca; patrick.brown@utoronto.ca; prabhat.jha@utoronto.ca; rashidansumana@gmail.com;

Abstract

Background: Verbal Autopsies (VAs) collect data on deaths and their causes outside of traditional hospital settings to provide more representative counts and Causes of Death (CODs) for reducing premature mortality. Current computer models for COD assignment in VAs perform similar to physicians at the population level, but poorly at the individual level, due to a focus on structured questionnaire data and neglecting free text from the open narratives. Recently, a Generative Pre-trained Transformer (GPT) model called ChatGPT-4 has demonstrated human-level performance on professional and academic exams using free

text input. ChatGPT-4 shows promise in mimicking physician behavior for assigning CODs, but to the best of our knowledge, has yet to be tested for assigning CODs using open narratives from VAs.
Methods: 6939 records collected from VA in Sierra Leone from 2019 to 2022 were used to compare four computer models, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA, to physicians for assigning CODs at the population and individual level. Open narratives were used for Chat-3.5/GPT-4 input, while structured questionnaires were used for InterVA-5/InSilicoVA input. All COD assignments were grouped into general COD categories consisting of 19, 10, and 7 categories for the adult, child, and neonatal age groups. Cause Specific Mortality Fraction (CSMF) accuracy and Partial Corrected Concordance (PCCC) were used to compare models to physicians at the population and individual level respectively. Comparisons in CSMF and PCCC to physicians among models were evaluated for all records and by COD, age group, and age ranges.
Results: x.
Conclusion: x.

Keywords: Cause of Death, Physician Coding, Verbal Autopsy, GPT

1 Background

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

VA involves two major components: survey and COD assignment [12–14]. In the survey component, trained lay surveyors interview those familiar with the deceased (e.g. living spouse, children, family, friends) to gather information using standardized questionnaires and open narratives. In the COD assignment component, physicians evaluate information available from the questionnaires and open narratives to assign probable CODs. Although VA surveys have been an effective alternative to collect mortality data at scale (e.g. less than $3 USD per house in India [15, 16]), COD assignment has been criticized to be expensive and difficult to reproduce due to reliance on physician assignment [17–19]. As an alternative to physician assignment, computer models, such as InterVA [20] and InSilicoVA [17], have recently been studied to automatically assign CODs with performances close to physicians at the population level, but poor performances at the individual level [21–25]. These computer models often utilized

data from the structured questionnaire, but often omit the free-text open narrative, which misses latent information, such as chronology or health-seeking behaviors, that may potentially help models perform better than using the questionnaire alone [26–28].

Recently, Large Language Models (LLM), leveraging massive datasets and deep learning approaches, have made advances in performing a variety of Natural Language Processing (NLP) tasks using free-text, such as question answering, code generation, and even medical diagnosis [29–32]. On November 30, 2022, a widely-available LLM called ChatGPT was released by OpenAI with capabilities of answering natural language text inquiries using training data up to September 2021. ChatGPT-3 was based on several Generative Pre-trained Transformer (GPT) models between 2018 to 2020, namely GPT-1 to GPT-3, which had notable differences in training data sizes of 5 gigabytes to 45 terabytes from web sources that resulted in 117 million to 175 billion parameter models [33]. On March 14, 2023, ChatGPT-4 was released with human-level performance on various professional and academic exams and benchmarks that outperformed ChatGPT-3 [34]. Given the limited usage of free-text open narratives in computer models for determining CODs, and recent advances in LLMs that leverage natural language text prompts, we conduct a case study with Sierra Leone deaths from VA in 2019 to 2022 to compare four models, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA, to physicians for determining CODs.

2 Methods

This study uses 6939 physician agreed records of 11,920 records collected in 2019 to 2022 from the Healthy Sierra Leone (HEAL-SL) study as described in Section 2.1. Section 2.2 describes the four models, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA, for COD assignment, and their inputs and outputs, while section 2.3 details the performance evaluation of the four models relative to physicians overall, by ages, and by COD, using population and individual level metrics. Figure A1 details the methods used in this study.

2.1 Verbal Autopsy (VA) Data

Initially, 11,920 records[1] from the HEAL-SL study [35, 36] were collected from dual-coded EVA, where each record was randomly coded by two different physicians that assigned CODs as International Classification of Diseases Revision 10 (ICD-10) codes [37]. Physicians were able to assign CODs for 11,820 of the 11,920 records, where 100 of these records could not be assigned a COD due to missing or inadequate information (e.g. low quality narrative, data loss). To determine if two codes were in physician agreement per record, codes were compared for similarity using Central Medical Evaluation Agreement 10 (CMEA-10) codes, which groups a range of similar ICD-10 codes together indicating codes in agreement [38] (see Additional File 2). When codes were not in agreement, a record enters the reconciliation phase, where the two physicians were provided reasoning and initial codes from each other to: (1) keep their initial code (2) assign the other physician's code or (3) assign a new code. If codes were not in agreement after the reconciliation phase, a record enters the adjudication phase,

[1] Available at openmortality.org/dataset/heal-sl

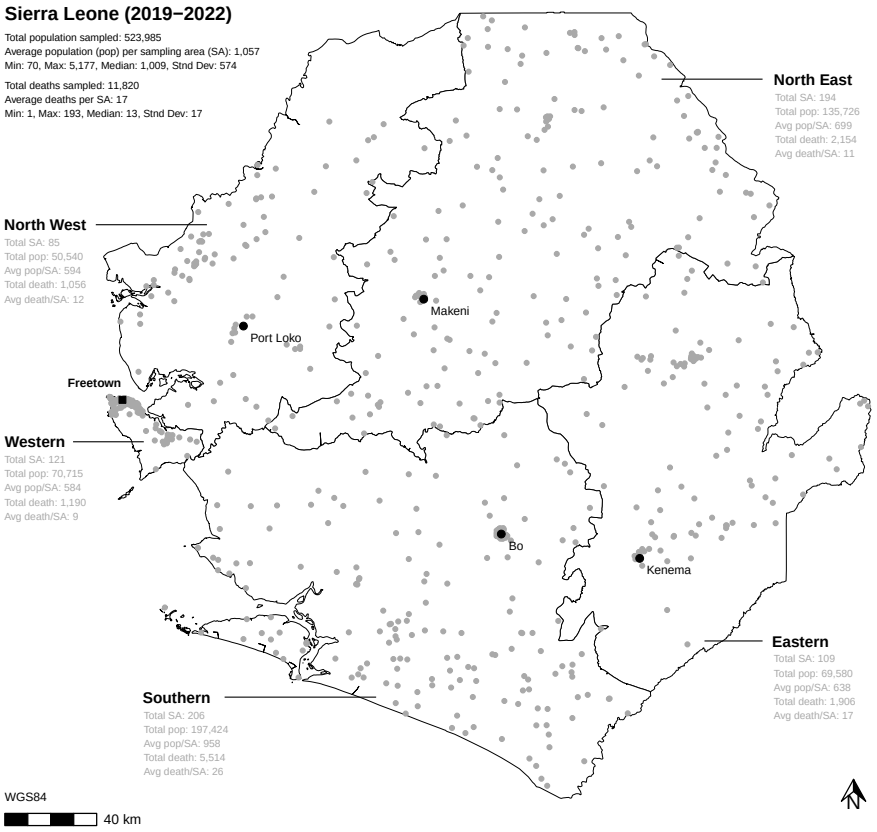

Fig. 1 Case study sample areas. VA records were collected from sample areas in Sierra Leone between 2019-2022 from the HEAL-SL study. Sample deaths shown represent records with physician assigned CODs only.

where a third senior physician evaluates both physicians' reasoning and codes before and after reconciliation, and assigns a final code based on their evaluation. The 11,820 physician coded records were further filtered for records where both physicians agreed on the assigned codes (records that were not reconciled or adjudicated) resulting in 6942 physician agreed records. Since computer models were compared to physicians in this study, there was more certainty that COD assignments agreed by both physicians were representative of physician assignment than when they disagreed [18, 39, 40]. 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 relatively evenly distributed (approximately 40-57% male and female) for the adult, child, and neonatal age groups in terms of sex (Figure B2) with the most records in the 1-5 year (n=1633, 24%) age range, and the least records in the 10-14 year (n=135, 2%) and 7-27 day (n=82, 1%) age ranges (Figure B3). The other 15 five-year age ranges within 0 days to 69 years were relatively evenly distributed with approximately 4-6% of the 6939 records each. The 3826 adult records had most records in the 65-69 year (n=575, 15%), and the least in the 10-14 year (n=135, 4%) age ranges, while the other 10 age ranges were relatively evenly distributed at 7-9% of all records (Figure C4). For adult CODs, Malaria (n=799, 21%) had the highest number of adult records, and cancers (n=49, 1%), diabetes mellitus (n=27, <1%), and suicide (n=3, <1%) had the lowest number adult of adult records, while the other 15 CODs had between 2-11% of all adult records (Figure C5). The 2636 child records had most records in the 1-5 year (n=1633, 62%) age range while the three other age ranges were relatively evenly distributed at 12-14% of all records (Figure D6). For child CODs, Malaria (n=1382, 52%) and other infections (n=667, 25%) had the highest number of child records, while nutritional deficiencies (n=11, <1%) and congenital anomalies (n=1, <1%) had the lowest number of records with the five other CODs between 3-7% of all child records. The 477 neonatal records had two age ranges with most records in the 0-6 day (n=395, 83%) age range and the rest in the 7-27 day (n=82, 17%) age range (Figure E8). For neonatal CODs, stillbirth (n=172, 36%) had the highest number of neonatal records, and ill-defined (n=23, 5%), other (n=5, 1%), and congenital anomalies (n=2, <1%) had the least number of records with the rest of the other three CODs between 15-22% of all neonatal records (Figure E9). See Appendices B to E for a list of figures B2 to E9 showing distributions of the 6939 physician agreed records.

2.2 Modelling

Four computer models were used to assign COD for each of the 6939 physician agreed records: GPT-3.5, GPT-4, InterVA-5, and InSilicoVA. Each model 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. All model outputs were converted to CGHR-10 codes to evaluate performances of models for COD assignment relative to physicians.

2.2.1 Overview of GPT-3.5/4, InterVA-5, and InSilicoVA

GPT-3.5 [41] and GPT-4 [34] are LLMs that utilize deep neural networks with transformer architectures [42] and reinforcement learning from human feedback [43–46] to follow instructions from prompts and provide human-level responses, with known differences in GPT-4 possessing multimodal capabilites (e.g. image/voice input/output), more recent training data, and improved responses compared to ChatGPT-3 [33]. GPT-3.5 and GPT-4 were instructed with text prompts to assign CODs based

on the open narrative. InterVA-5 and InSilicoVA are widely used and studied standard statistical models [13, 21, 22, 24, 25, 47, 48] for COD assignment in VAs under the openVA framework [49]. InterVA-5 applies Bayesian probabilistic modelling [50] using a set of standardized symptoms from reports and related conditional probabilities from medical experts to assign CODs based on the highest probability [20, 51]. InSilicoVA improves upon InterVA (e.g. comparable probabilities across individuals, measures of uncertainty, and inclusion of additional data sources) with a hierarchical Bayesian framework and Markov Chain Monte Carlo (MCMC) simulations [52–54] to incorporate multiple sources of uncertainty for assigning CODs based on the highest probability [17]. For assigning CODs, GPT-3.5 and GPT-4 require prompts containing conversation-like textual instructions as input, while InterVA-5 and InSilicoVA require structured symptom and associated sociodemographic data as input.

2.2.2 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 expertisein 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 [55, 56] consisting of 526 variables [57], followed by further conversion into OpenVA format [49] consisting of 353 variables [58] using the `pyCrossVA` version 0.97 Python package [59]. The 6939 records were all converted into OpenVA formatted records for InterVA-5 and InSilicoVA.

2.2.3 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 [60]. 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 [61] for InSilicoVA was set to 9500[2], while the `HIV` (level of prevalence of human immunodeficiency virus) and `Malaria` (level of prevalence of Malaria) parameters [62] for InterVA-5 were both set to `'h'` (high) reflecting HIV and Malaria disease assumptions in Sierra Leone [63, 64].

2.2.4 Output Data and Code Conversions

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 [65] 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.

2.3 Performance Evaluation

The performance of four models, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA, were evaluated with metrics on the population and individual level by comparing their CGHR-10 COD outputs to 6939 records. Cause Specific Mortality Fraction (CSMF) accuracy was used to evaluate models on the population level, while Partial Chance Corrected Concordance (PCCC) was used to evaluate models on the individual level [66]. 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 by three age groups, and further into age ranges and COD for each age group.

2.3.1 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 [66]. To calculate CSMF accuracy, we first calculate $CSMF_j$ 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 1. Then, the $CSMFMaximumError$, representing the worst possible model, is calculated using Equation 2. Finally, the

[2]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.

CSMF accuracy is given by Equation 3, 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.

$$CSMF_j = Records_j / Records \tag{1}$$

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

$$CSMFAccuracy = 1 - \frac{\sum_{j=1}^{k} |CSMF_j^{true} - CSMF_j^{pred}|}{CSMFMaximumError} \tag{3}$$

2.3.2 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 [66]. PCCC is given by Equation 5, 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, we set k 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 4. 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.

$$C = \frac{TP}{Records} \tag{4}$$

$$PCCC(k) = \frac{C - \frac{k}{N}}{1 - \frac{k}{N}} \tag{5}$$

2.3.3 Metrics Overall and by Age and Cause of Death

The CSMF accuracy and PCCC metrics were calculated and compared for each model overall and by age group, followed by age ranges and COD for each age group as model performance can vary across ages and specific causes [47, 48, 67]. Metrics were calculated overall for three age groups according to the CGHR-10 codes: adult (12 to 69 years), child (28 days to 11 years), and neonatal (under 28 days). For each of the adult and child age groups, metrics were calculated for five-year age ranges for records with ages at death of one-year or older and five-month age ranges for 28 days or older. For the neonatal age group, the age ranges of 0-6 days and 7-27 days were used. Metrics were also calculated by CODs defined by CGHR-10 codes, which include 19, 10, and 7 CODs for adult, child, and neonatal records respectively.

3 Results

This section details the performance results of GPT-3.5, GPT-4, InterVA-5, and InSilicoVA models for assigning CGHR-10 CODs after applying the methods in Section 2. GPT-4 performed the best overall at 0.71 PCCC followed by GPT-3.5 at 0.64 PCCC. GPT-4 also had the highest PCCC for most age ranges and CODs across the adult (12 to 69 years), child (28 days to 11 years), and neonatal (under 28 days) age groups with GPT-3.5, InterVA-5, and InSilicoVA having higher PCCC values for some age ranges and CODs. Overall performance results are seen in Section 3.1, and performance by adult, child, and neonatal records are seen in Sections 3.2, 3.3, and 3.4 respectively.

3.1 Overall Performance

Of the 6939 records, GPT-4 had the highest PCCC followed by GPT-3.5, InSilicoVA, and InterVA-5 at 0.61, 0.58, 0.44, and 0.43 PCCC respectively (Table 1). 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 0.03 PCCC. CSMF accuracies were above 0.7 for all models at 0.79, 0.74, 0.74, and 0.78 for GPT-4, GPT-3.5, InSilicoVA, and InterVA-5 respectively. Figure 2 shows the PCCC performance across age groups with reference to the CSMF Accuracy. GPT-4 had the highest PCCC at 0.65 and 0.62 for the adult and neonatal age groups, while GPT-3.5 had the highest PCCC at 0.57 for the child age group with GPT-4 performing slightly worse at 0.54. GPT-3.5 had the second highest PCCC at 0.56 fo the adult age group, while InSilicoVA had the second highest PCCC for the neonatal age group at 0.56. InSilicoVA and InterVA-5 performed the worse for both adult and child age groups at PCCC scores below 0.5, while GPT-3.5 performed the worse for the neonatal age group with a PCCC of 0.42. The PCCC for models had larger differences between the minimum and maximum PCCC for the adult age group at 0.24, than the child and neonatal age groups at 0.12 and 0.20 respectively. For reference, all records, including non-physician agreed records, were also coded by GPT-3.5, GPT-4, InterVA-5, and InSilicoVA for comparison to physician agreed records. For physician agreed records, models had 0.05 to 0.09 higher PCCC values and 0.01 to 0.05 higher CSMF accuracy values than for all records (see Figure F10 in Appendix F).

Table 1 Overall model performance for physician agreed records (n=6939)

Model	PCCC	CSMF Accuracy	PCCC Improvement
GPT-4	0.61	0.79	↑ 0.03
GPT-3.5	0.58	0.74	↑ 0.14
InSilicoVA	0.44	0.74	↑ 0.01
InterVA-5	0.43	0.78	-

Note: Models sorted from highest to lowest PCCC. PCCC improvements are from model below.

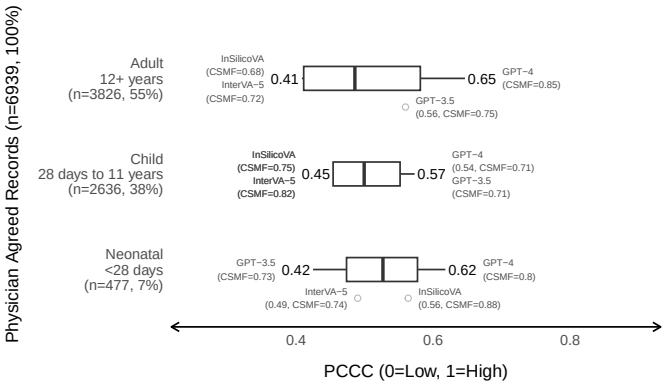

Fig. 2 Model performance by age group. GPT-3.5, GPT-4, InterVA-5, and InSilicoVA model performance for assigning CGHR-10 CODs in the adult (12 to 69 Years), child (28 days to 11 years), and neonatal (under 28 days) age groups. CSMF indicates CSMF accuracy.

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

Across the adult age ranges, GPT-4 had the highest PCCC values at or above 0.59, followed by GPT-3.5, while InterVA-5 and InSilicoVA had the lowest PCCC values at or below 0.46. GPT-4 PCCC was between 0.61 and 0.72 for young to middle age ranges between 10 to 49 years, and at or below 0.62 for older age ranges between 50 to 69 years (Figure 3). Each age range had between 135 (4%) to 575 (15%) records with the 65-69 year age range having the most records (n=575, 15%), while other age ranges ranged from 4-9% of all records. Figure 4 shows the highest and lowest performing models by PCCC across 18 adult CODs. The highest PCCC values were GPT-4 between 0.36 to 0.99 for 12 of 18 CODs (GPT-3.5 within 0.05 for 7 of the 15 CODs), InSilicoVA for road and transport injuries at 0.84 (InterVA-5 close at 0.81), GPT-3.5 for ill-defined at 0.57,and InterVA-5 for ischemic heart disease at 0.61 (GPT-4 close at 0.57). The lowest PCCC values were InterVA-5 between 0 to 0.79 for 11 of 18 CODs (InSilicoVA within 0.05 for 3 of the 11 CODs), InSilicoVA between 0.01 to 0.41 for 6 of 18 CODs (InterVA-5 within 0.05 for other cardiovascular diseases and diarrhoeal diseases), and GPT-3.5 for ischemic heart disease at 0.49 (InSilicoVA within 0.05) and road and transport injuries at 0.59. Suicide and chronic respiratory diseases had the largest differences between minimum and maximum PCCC at 1 and 0.81 respectively. Maternal conditions, road and transport injuries, and tuberculosis had higher PCCC values above or at 0.67 (with the exception of GPT-3.5 for road and transportation injuries as an outlier), where there were small differences between minimum and maximum PCCC ranging from 0.12 to 0.19. Ischemic heart disease and unspecified infections had lower PCCC values below or at 0.6, where there were small differences between minimum and maximum PCCC ranging at 0.02 and 0.17 respectively. Each adult COD had between 3 (<1%) to 799 records (21%) with malaria (n=799, 21%), other noncommunicable diseases (n=414, 11%), and unspecified infections (n=389, 10%) making 42% of all records, suicide having less than 1% of records, and other CODs between 1-9% of all records. GPT-4 had the highest PCCC for both male and female records at 0.72 and 0.7 respectively, while all other models had PCCC at or below 0.63 (see Figure F11 in Appendix F).

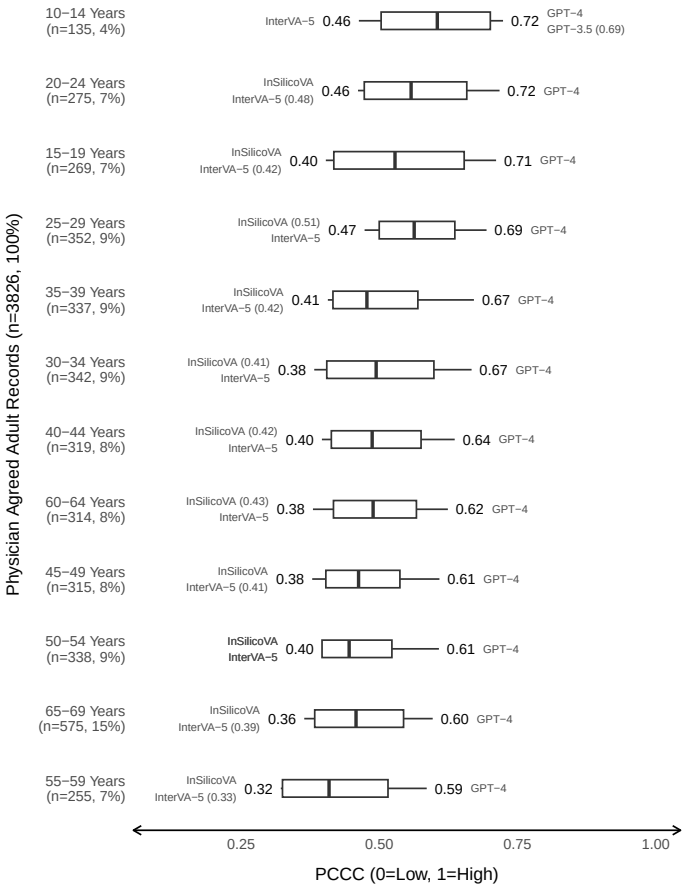

Fig. 3 Model performance for adult records by age range. GPT-3.5, GPT-4, InterVA-5, and InSilicoVA model performance for assigning CGHR-10 CODs in the adult age group (12 to 69 Years) by five-year age ranges. PCCC values are sorted by the highest to lowest PCCC for each age range. Only models with PCCC values within .05 of the highest and lowest PCCC are shown.

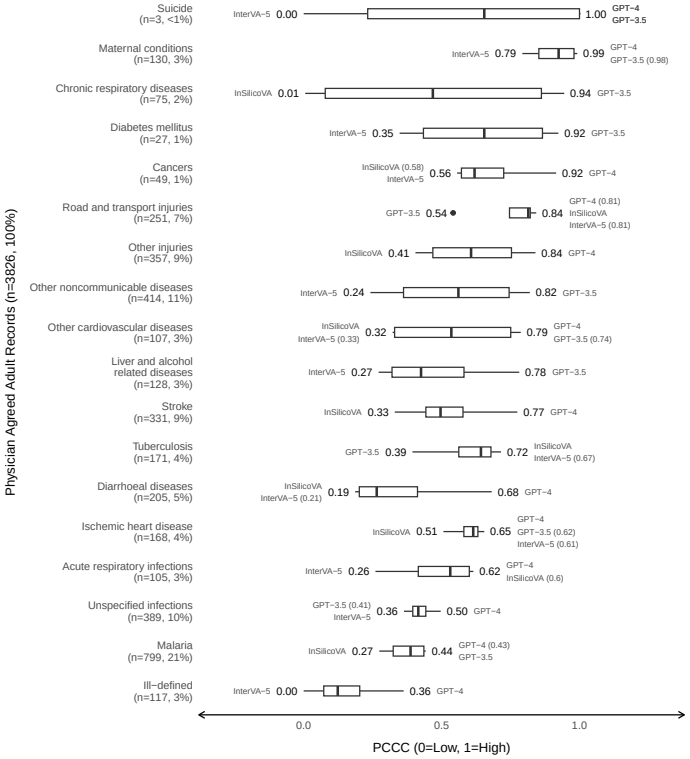

Fig. 4 Model performance for adult records by COD. GPT-3.5, GPT-4, InterVA-5, and InSilicoVA model performance for assigning CGHR-10 CODs in the adult age group (12 to 69 Years) by COD. PCCC values are sorted by the highest to lowest PCCC for each COD. Only models with PCCC values within .05 of the highest and lowest PCCC are shown.

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

Across the child age ranges, GPT had the highest PCCC at 0.69 to 0.75 for all age ranges (GPT-3.5 within 0.05 for 1-11 years), while the lowest PCCC values were for InSilicoVA for 1-11 years (InterVA-5 within 0.05 for 1-5 years), and InterVA-5 for 1-11 months (Figure 5). Each child age range had between 308 (12%) to 1633 (62%) records with most records in the 1-5 years range at 1633 records (62%) and the other age ranges consisting of 12-14% of all records. Figure 6 shows the highest and lowest performing models by PCCC across 9 child CODs. GPT-4 had the highest PCCC for 7 of 9 child CODs, where GPT-3.5 had 5, InterVA-5 had two (congenital anomalies and diarrhoeal diseases), and InSilicoVA had one (pneumonia) of the 7 CODs within 0.05 PCCC of GPT-4. InterVA-5 had the highest PCCC for diarrhoeal disease at 0.82 (GPT-4 at 0.81), and GPT-3.5 had the highest PCCC for ill-defined at 0.84. InSilicoVA had the lowest PCCC for 4 of 9 child CODs (InterVA-5 within 0.05 for 3 of the 4 CODs), InterVA-5 had the lowest PCC for the other 4 of 9 child CODs (InSilicoVA within 0.05 for injuries), and GPT-3.5 had the lowest PCCC for diarrhoeal diseases at 0.69. Congenital anomalies had one record only in which GPT-4, GPT-3.5, and InterVA-5 assigned a CGHR-10 code matching the physicians. Ill-defined had the largest difference between minimum and maximum PCCC at 0.84, while injuries, diarrhoeal disease, and pneumonia (with the exception of InterVA-5 as an outlier at 0.41) had higher PCCC values at or above 0.69 with small differences between minimum and maximum PCCC ranging from 0.01 to 0.14. Other infections had lower PCCC values at or below 0.54 with a small difference between minimum and maximum PCCC at 0.2. Each child COD had between 1 (<1%) to 1382 (52%) records with malaria (n=1382, 52%), other infections (n=667, 25%), and diarrhoeal diseases (n=175, 7%) making 84% of all records. Congenital anomalies and nutritional deficiencies had less than 1% of all records, while other CODs had between 1-3% of all records. GPT-4 had the highest PCCC for both male and female records at 0.73 and 0.72 (GPT-3.5 at 0.69) respectively, while all other models had PCCC at or below 0.68 (see Figure F12 in Appendix F).

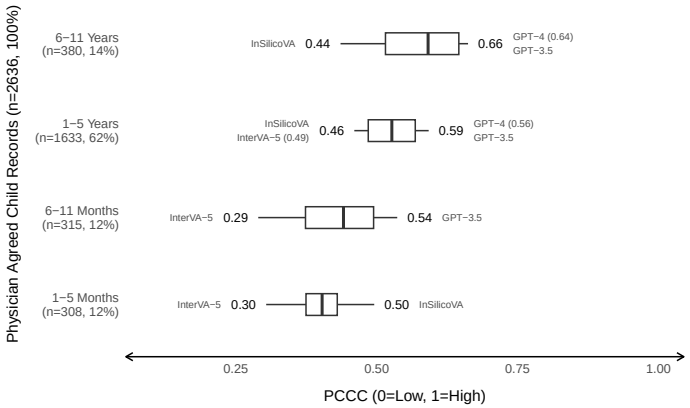

Fig. 5 Model performance for child records by age range. GPT-3.5, GPT-4, InterVA-5, and InSilicoVA model performance for assigning CGHR-10 CODs in the child age group (28 days to 11 years) by five-year age ranges. PCCC values are sorted by the highest to lowest PCCC for each age range. Only models with PCCC values within .05 of the highest and lowest PCCC are shown.

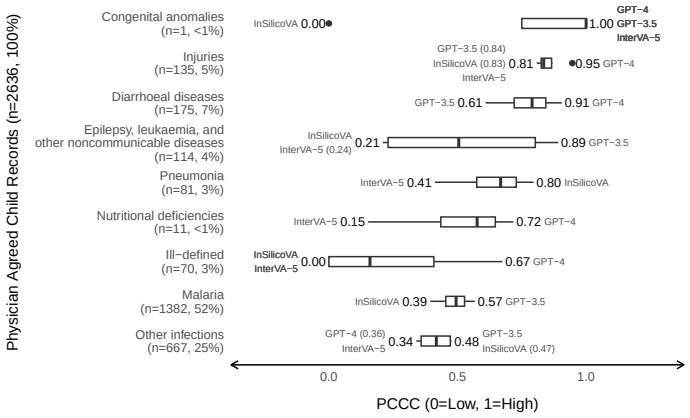

Fig. 6 Model performance for child records by COD. GPT-3.5, GPT-4, InterVA-5, and InSilicoVA model performance for assigning CGHR-10 CODs in the child age group (28 days to 11 years) by COD. PCCC values are sorted by the highest to lowest PCCC for each COD. Only models with PCCC values within .05 of the highest and lowest PCCC are shown.

3.4 Performance for 477 Neonatal Records (Under 28 Days)

GPT-4 and InSilicoVA had the highest PCCC for 0-6 day and 7-27 day neonatal age ranges at 0.62 and 0.67 respectively, while InterVA-5 had the lowest PCCC across all neonatal age ranges (GPT-3.5 within 0.05) as seen in Figure 7. There were small differences between minimum and maximum PCCC across all age ranges at 0.14 and 0.12 for 0-6 days and 7-27 days respectively. Most records were in the 0-6 day age range at 395 (83%) records, while the rest were in the 7-27 day age range at 82 (17%) records. Figure 8 shows the performance of models by PCCC across 7 neonatal CODs. GPT-4 had the highest PCCC for 4 of the 7 CODs (GPT-3.5 identical to GPT-4 for congenital anomalies at 1 PCCC). InSilicoVA, GPT-3.5, and InterVA-5 had the highest PCCC for neonatal infections at 0.87, ill-defined at 0.59 (GPT-4 at 0.55), and birth asphyxia and birth trauma at 0.59 respectively. Congenital anomalies and other had less or equal to 5 records with large difference of 1 between minimum and maximum PCCC. Neonatal infections and stillbirth had smaller differences between minimum and maximum PCCC at 0.38 and 0.24 respectively with larger PCCC values at or above 0.49. Birth asphyxia and birth trauma also had smaller differences between minimum and maximum PCCC at 0.24 but had PCCC values at or below 0.59. Prematurity and low birthweight and ill-defined had large differences between minimum and maximum PCCC at 0.62 and 0.58 respectively. Each neonatal COD had between 2 (<1%) to 172 (38%) records, with stillbirth (n=172, 36%), birth asphyxia and birth trauma (n=103, 22%), and neonatal infections (n=99, 21%), and prematurity and low birthweight (n=73, 15%) making 94% of all records, while the rest are in congenital anomalies, other, and ill-defined CODs. GPT-4 had the highest PCCC for both female and male records at 0.65 and 0.58 (InSilicoVA at 0.56) respectively, while all other models had PCCC at or below 0.56 (see Figure F13 in Appendix F).

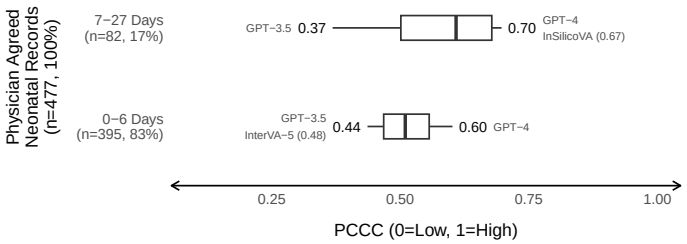

Fig. 7 Model performance for neonatal records by age range. GPT-3.5, GPT-4, InterVA-5, and InSilicoVA model performance for assigning CGHR-10 CODs in the neonatal age group (under 28 days) by five-year age ranges. PCCC values are sorted by the highest to lowest PCCC for each age range. Only models with PCCC values within .05 of the highest and lowest PCCC are shown.

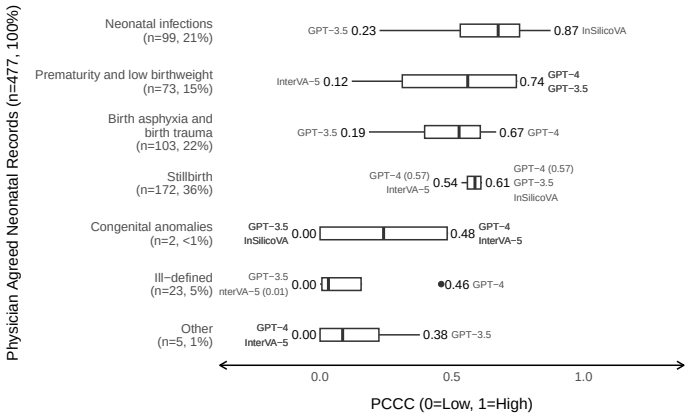

Fig. 8 Model performance for neonatal records by COD. GPT-3.5, GPT-4, InterVA-5, and InSilicoVA model performance for assigning CGHR-10 CODs in the neonatal age group (under 28 years) by COD. PCCC values are sorted by the highest to lowest PCCC for each COD. Only models with PCCC values within .05 of the highest and lowest PCCC are shown.

4 Discussion

TO DO.

4.1 Advantages

x.

Captured causes of death with very low number of records.

4.2 Disadvantages

x.

Dimensions, reliabiity, and cost of LLM versus physician.

Reliance on third-party non-open source models.

4.3 Limitations

x.

Limited records for some causes of death. Did not explore adjudicated and reconciliated records.

Did not explore effects of model parameters outside the default (e.g. ChatGPT temperature, etc).

Did not explore response consistency with reruns.

4.4 Opportunities

x.

Prompt engineering: adjust and optimize prompts, add in correctional instructions, adjudication, reconciliation, ICD-10/physician coding guide standards, and system prompt.

Sensitivity analysis: Test consistency of GPT responses with multiple reruns and parameter settings.

5 Conclusion

TO DO.

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.

Declarations

Availability of data and materials

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

Appendix A Case Study Methods

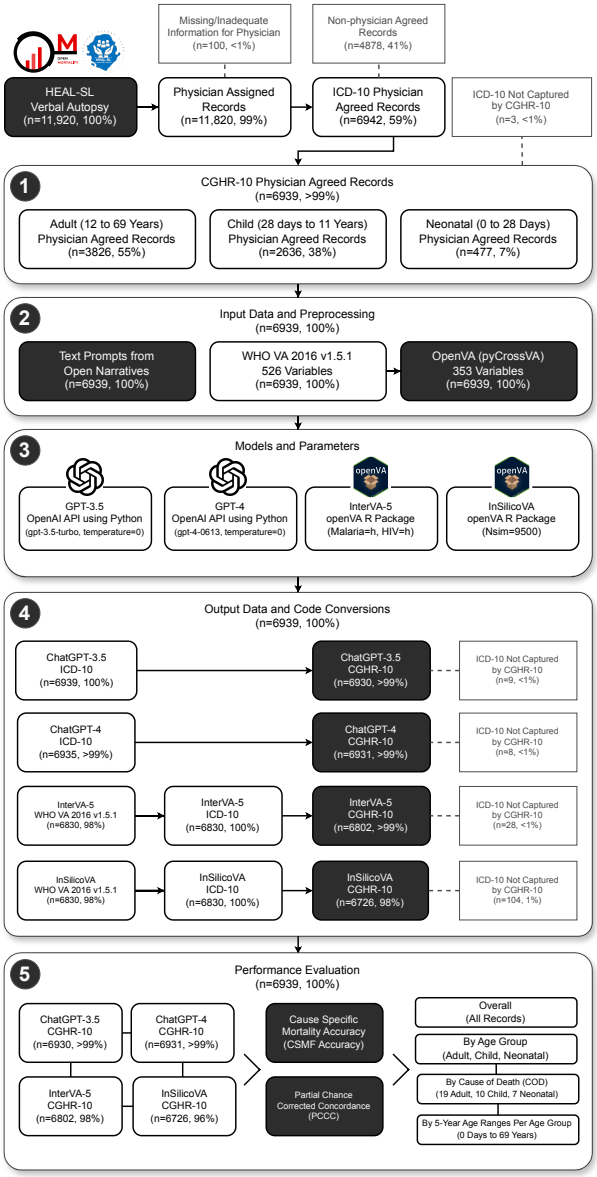

Fig. A1 Case study methods. VA records in Sierra Leone between 2019-2022 from the HEAL-SL study were filtered for physician agreed records, and pre-processed for COD assignment by GPT-3.5, GPT-4, InterVA-5 and InSilicoVA models. The output models were evaluated with CSMF Accuracy and PCCC overall, by age group, and by COD and five-year age ranges for each age group.

Appendix B Physician Agreed Verbal Autopsy Records

The following figures visualize the distribution of all physician agreed records by age groups (Figure B2) and 5-year age ranges (Figure B3) in Sierra Leone from 2019 to 2022.

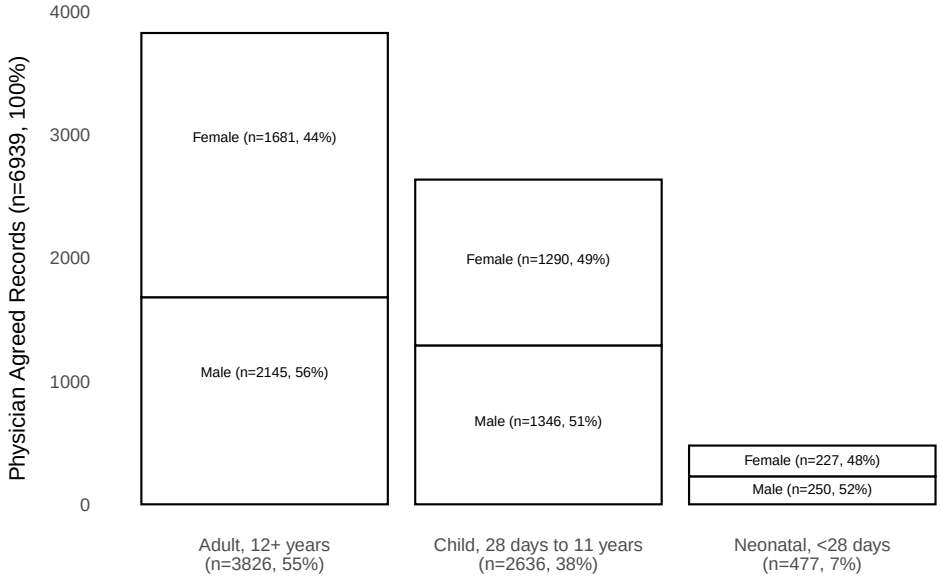

Fig. B2 Number of physician agreed VA records by adult, child, and neonatal age groups in Sierra Leone from 2019 to 2022.

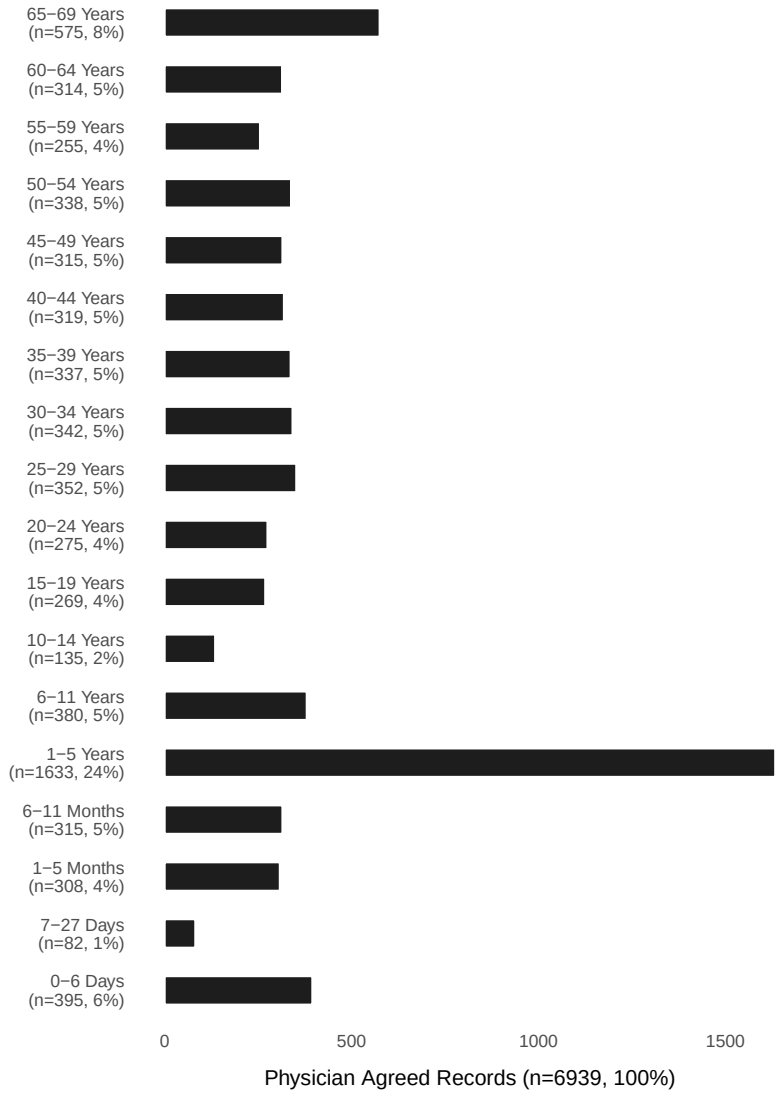

Fig. B3 Number of physician agreed VA records by 5-year age ranges in Sierra Leone from 2019 to 2022.

Appendix C Adult Physician Agreed Verbal Autopsy Records (12 to 69 Years)

The following figures visualize the distribution of adult physician agreed records by 5-year age ranges (Figure C4) and CGHR-10 COD code in Sierra Leone from 2019 to 2022 (Figure C5).

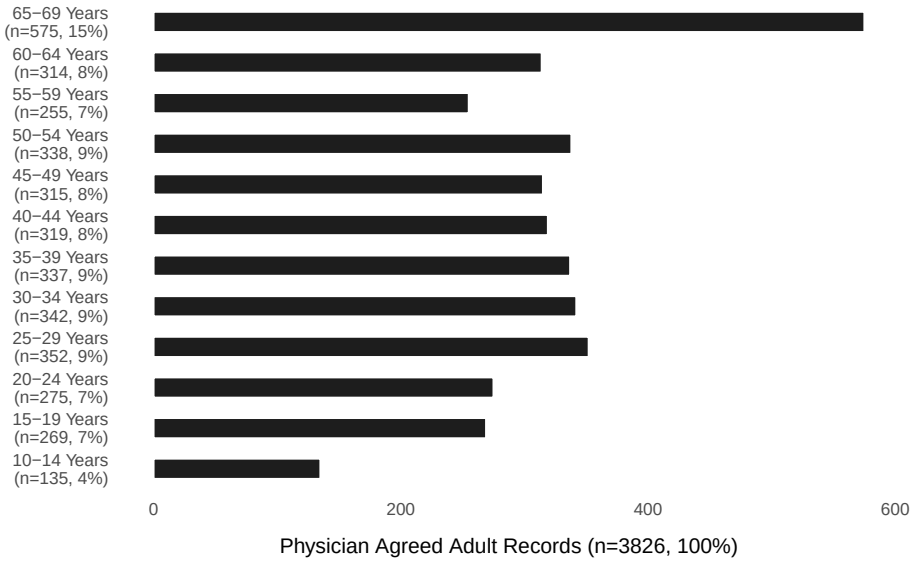

Fig. C4 Number of physician agreed adult (12 to 69 years) VA records by 5-year age ranges in Sierra Leone from 2019 to 2022.

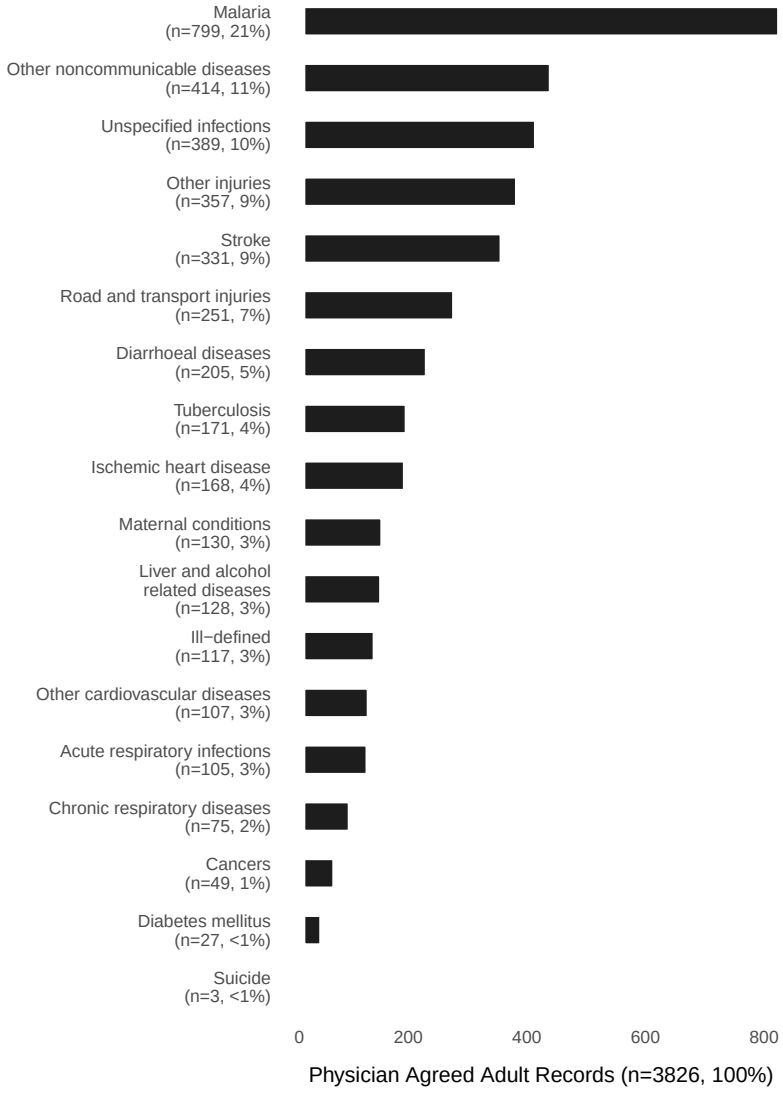

Fig. C5 Number of physician agreed adult (12 to 69 years) VA records by CGHR-10 COD code in Sierra Leone from 2019 to 2022.

Appendix D Child Physician Agreed Verbal Autopsy Records (28 Days to 11 Years)

The following figures visualize the distribution of child physician agreed records by 5-year age ranges (Figure D6) and CGHR-10 COD code in Sierra Leone from 2019 to 2022 (Figure D7).

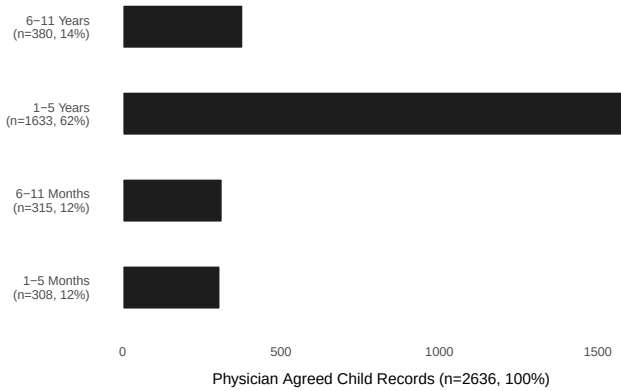

Fig. D6 Number of physician agreed child (28 days to 11 years) VA records by 5-year age ranges in Sierra Leone from 2019 to 2022.

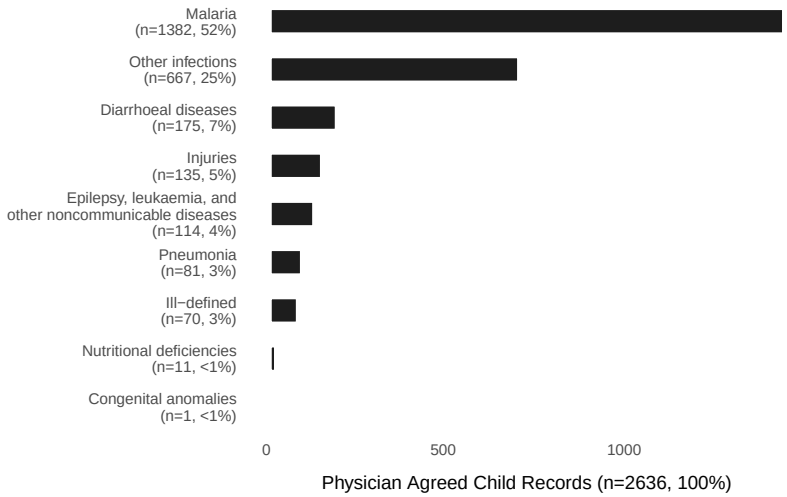

Fig. D7 Number of physician agreed child (28 days to 11 years) VA records by CGHR-10 COD code in Sierra Leone from 2019 to 2022.

Appendix E Neonatal Physician Agreed Verbal Autopsy Records (Under 28 Days)

The following figures visualize the distribution of neonatal physician agreed records by 5-year age ranges (Figure E8) and CGHR-10 COD code in Sierra Leone from 2019 to 2022 (Figure E9).

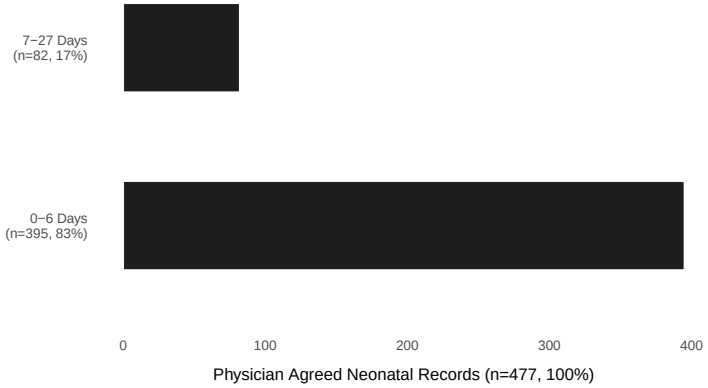

Fig. E8 Number of physician agreed neonatal (less than 28 days) VA records by 5-year age ranges in Sierra Leone from 2019 to 2022.

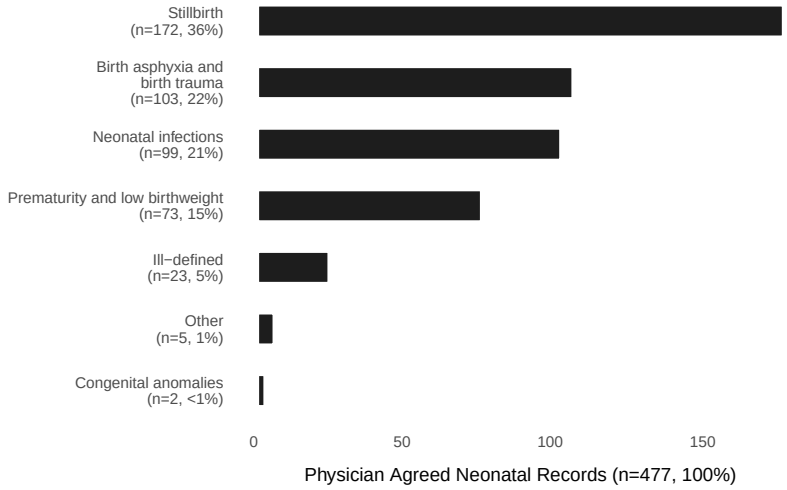

Fig. E9 Number of physician agreed neonatal (less than 28 days) VA records by CGHR-10 COD code in Sierra Leone from 2019 to 2022.

Appendix F Performance By Other Variables

The following figures show the model performance by PCCC for all records and physician agreed records (Figure F10), and the performance of models by PCCC across sex for the adult (12 to 69 years), child (28 days to 11 years), and neonatal (under 28 days) age groups in Figures F11, F12, and F13 respectively.

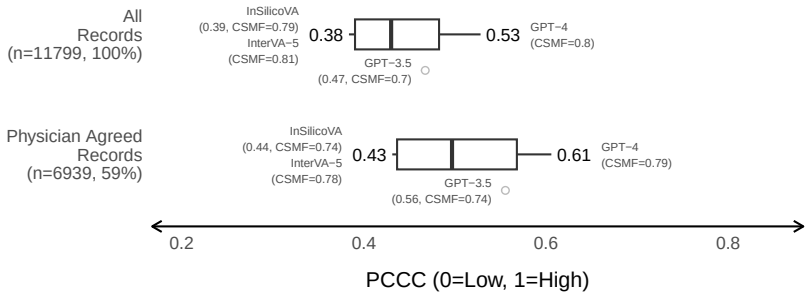

Fig. F10 Model performance for all records versus physician agreed records. GPT-3.5, GPT-4, InterVA-5, and InSilicoVA model performance for assigning CGHR-10 CODs for all records and physician agreed records. PCCC values are sorted by the highest to lowest PCCC.

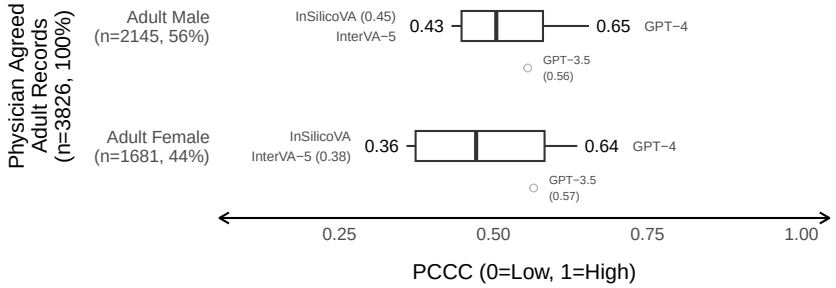

Fig. F11 Model performance for adult records by sex. GPT-3.5, GPT-4, InterVA-5, and InSilicoVA model performance for assigning CGHR-10 CODs in the adult age group (12 to 69 years) by sex. PCCC values are sorted by the highest to lowest PCCC.

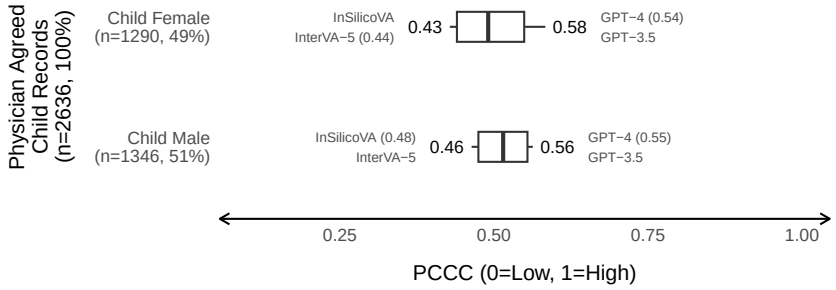

Fig. F12 Model performance for child records by sex. GPT-3.5, GPT-4, InterVA-5, and InSilicoVA model performance for assigning CGHR-10 CODs in the child age group (28 days to 11 years) by sex. PCCC values are sorted by the highest to lowest PCCC.

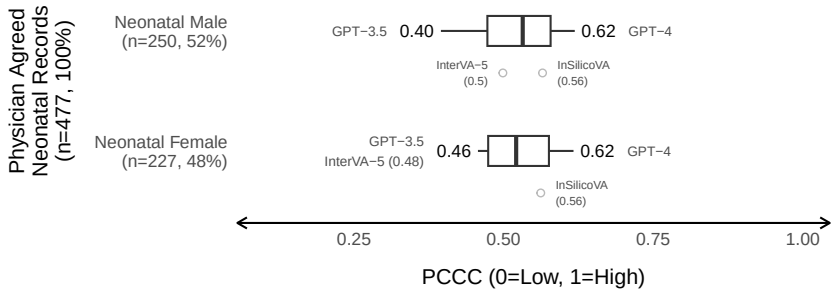

Fig. F13 Model performance for neonatal records by sex. GPT-3.5, GPT-4, InterVA-5, and InSilicoVA model performance for assigning CGHR-10 CODs in the neonatal age group (under 28 days) by sex. PCCC values are sorted by the highest to lowest PCCC.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

[14] World Health Organization. Verbal Autopsy Standards: Ascertaining and Attributing Cause of Death. World Health Organization; 2007. Available from: https://apps.who.int/iris/bitstream/handle/10665/43764/9789241547215_eng.pdf.

[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. Available from: http://arxiv.org/abs/2307.03109.

[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. Available from: https://doi.org/10.1145/3368089.3417058.

[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. Available from: http://arxiv.org/abs/2303.08774.

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

[42] 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. Available from: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[43] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al.: Training Language Models to Follow Instructions with Human Feedback. arXiv. Available from: http://arxiv.org/abs/2203.02155.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

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

[51] 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.

[52] 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.

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

[54] 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.

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

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

[57] DiPasquale A, Maire N, Bratschi M.: Release ODK 2016 WHO VA Instrument 1.5.1 SwissTPH/WHO-VA. Swiss Tropical and Public Health Institute. Available from: https://github.com/SwissTPH/WHO-VA/releases/tag/1.5%2C1.

[58] Byass P.: InterVA-5.1 User Guide. Available from: http://www.byass.uk/interva/products.

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

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

[61] Li ZR, McCormick T, Clark S.: InSilicoVA: Probabilistic Verbal Autopsy Coding with 'InSilicoVA' Algorithm. Available from: https://cran.r-project.org/package=InSilicoVA.

[62] Thomas J, Li Z, Byass P, McCormick T, Boyas M, Clark S.: InterVA5: Replicate and Analyse 'InterVA5'. Available from: https://cran.r-project.org/package=InterVA5.

[63] 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).

[64] 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.

[65] World Health Organization.: Verbal Autopsy Standards: The 2016 WHO Verbal Autopsy Instrument. Available from: https://www.who.int/publications/m/

item/verbal-autopsy-standards-the-2016-who-verbal-autopsy-instrument.

[66] 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.

[67] 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.