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], Asha Behdinan[1], Ronald Carshon-Marsh[1], Thomas Kai Sze Ng[1], Rashid Ansumana[2], Patrick Brown[1], Prabhat Jha[1]

[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; kaisze.ng@unityhealth.to; rashidansumana@gmail.com; patrick.brown@utoronto.ca; prabhat.jha@utoronto.ca;

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 GPT-3.5/4 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: Based on PCCC, GPT-4 had the best performance overall (0.61), followed by GPT-3.5 (0.58), InSilicoVA (0.44), and InterVA-5 (0.43). GPT-4 also had the best performance for the adult (0.65), child (0.57), and neonatal (0.62) age groups, while performance decreased as adults aged (0.72-0.59) and increased as children and neonates aged (0.5-0.7). 7 of 19, 4 of 10, and 1 of 7 adult, child, and neonatal CODs respectively had models with $\geq$0.8 PCCC. For most CODs, GPT-4 models had one of the highest performances with the exception of InSilicoVA for adult tuberculosis, child pneumonia and neonatal infections. At the population level, all models had CSMF accuracy between 0.7-0.79.
Conclusion: GPT and InSilicoVA models were comparable to physicians for some CODs, but require further evaluation on reliability and larger datasets.

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 [**?**]. 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 [**?**]. Thus, reliable counts and diagnoses of deaths enable decision makers to identify populations at risk to save lives and reduce premature deaths worldwide [**? ? ? ?**]. 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 [**?**]. 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 [**? ? ?**], outside of traditional healthcare facilities where over half of deaths occur at home [**?**].

VA involves two major components: survey and COD assignment [**? ? ?**]. 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. This component has been criticized to be difficult to reproduce due to reliance on physician assignment [**? ? ? ? ?**]. As an alternative to physician assignment, computer models, such as InterVA [**?**] and InSilicoVA [**?**], have been studied to automatically assign CODs with performances close to physicians at the population level, but poor performances at the individual level [**? ? ? ? ?**]. These computer models often utilize 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 [**? ? ?**].

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 [**? ? ? ?**]. In 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 [**?**]. In March 2023, ChatGPT-4 was released with human-level performance on various professional and academic exams and benchmarks that outperformed ChatGPT-3 [**?**]. 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 (two physicians with similar COD assignment on the same record) of 11,920 records collected in 2019 to 2022 from the Healthy Sierra Leone (HEAL-SL) study as described in Section **??**. Section **??** describes the four models, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA, for COD assignment, and their inputs and outputs, while section **??** details the performance evaluation of the four models relative to physicians overall, by ages, and by COD, using population and individual level metrics. See Appendix **??** for additional details on the methods used in this study.

2.1 Verbal Autopsy (VA) Data

Initially, 11,920 records from the HEAL-SL study [**?** **?**] 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 [**?**]. 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). For each record, two codes were assigned

by two different randomly selected physicians, where codes were evaluated for agreement using Central Medical Evaluation Agreement 10 (CMEA-10) codes. CMEA-10 groups a range of similar ICD-10 codes together, where if they are in agreement if they are within the same group [**?**] (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, 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 [**? ? ?**]. 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 evenly distributed in terms of sex and age, except for children aged 1-5 years (n=1633, 24%), while Malaria (n=799, 21%) was the COD for about a fifth of the records. See Appendix **??** for further details on data distributions across age groups.

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 [**?**] and GPT-4 [**?**] are LLMs that utilize deep neural networks with transformer architectures [**?**] and reinforcement learning from human feedback [**? ? ? ?**] 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 [**?**]. 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 [**? ? ? ? ? ? ?**] for COD assignment in VAs under the openVA framework [**?**]. InterVA-5 applies Bayesian probabilistic modelling [**?**] using a set of standardized symptoms from reports and related conditional probabilities from medical experts to assign CODs based on the highest probability [**? ?**]. 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 [**? ? ?**] to incorporate multiple sources

of uncertainty for assigning CODs based on the highest probability [**?**]. 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 [**? ?**] consisting of 526 variables [**?**], followed by further conversion into OpenVA format [**?**] consisting of 353 variables

[**?**] using the `pyCrossVA` version 0.97 Python package [**?**]. 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 [**?**]. 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 [**?**] for InSilicoVA was set to 9500, while the `HIV` (level of prevalence of human immunodeficiency virus) and `Malaria` (level of prevalence of Malaria) parameters [**?**] for InterVA-5 were both set to `'h'` (high) reflecting HIV and Malaria disease assumptions in Sierra Leone [**? ?**].

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 [**?**] 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 (see Appendix **??**), while Partial Chance Corrected Concordance (PCCC) was used to evaluate models on the individual level (see Appendix **??**) [**?**]. 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.

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 [**? ? ?**]. 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 **??**. GPT-4 performed the best overall at 0.61 PCCC followed by GPT-3.5 at 0.56 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 a few age ranges and CODs. Overall performance results are seen in Section **??**, and performance by adult, child, and neonatal records are seen in Sections **??**, **??**, and **??** respectively.

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 **??**). 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 **??** 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). Across age ranges, all models followed a similar pattern in individual performance (Figure **??**). PCCC rose for records with ages from 0 to 27 days, dropped drastically for 1-5 months, rose again from 6 months to 14 years, and slowly declined for ages 15 to 69 years. The highest and lowest performances were observed for ages 10-29 years and 1-11 months respectively. Performances varied more across models for ages 0 days to 5 years, while less variation

was seen between GPT-3.5 and GPT-4, as well as InSilicoVA and InterVA-5, from 5 to 69 years.

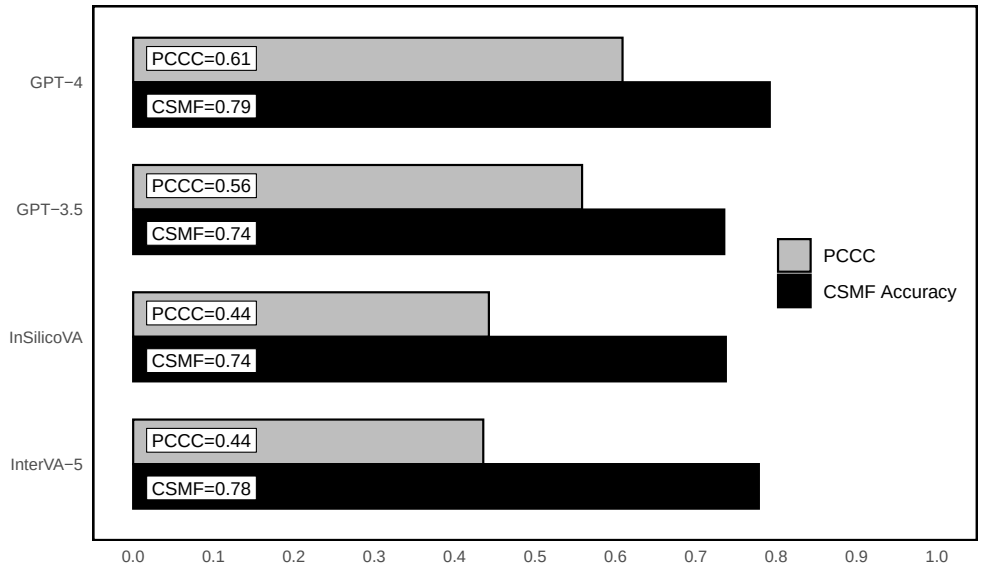

Fig. 1 Overall model performance.

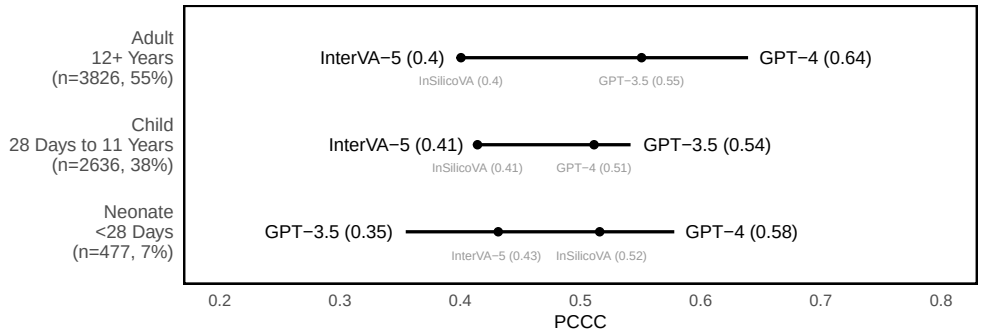

Fig. 2 Model performance by age group.

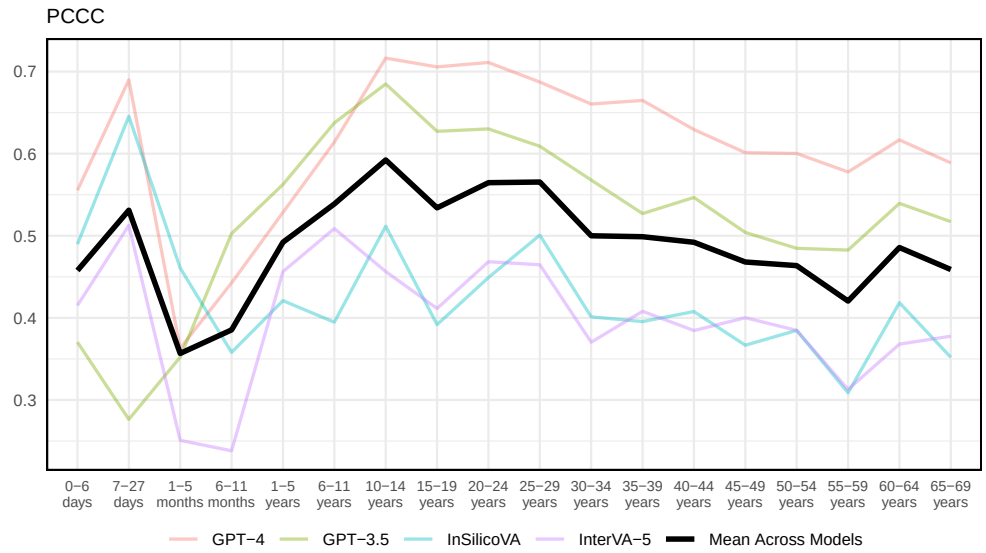

Fig. 3 Model performance by age range.

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

Figure **??** shows model performance by PCCC across 17 adult CODs excluding suicide due to low sample size (n=3, <1%). GPT-4 had the highest individual performance for 10 of 17 CODs (0.35 to 0.99 PCCC), GPT-3.5 for 5 CODs (0.43-0.94 PCCC), and InSilicoVA for 2 CODs (0.71 and 0.84 PCCC). InterVA-5 had the lowest performance for 8 of 17 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). GPT-3.5/4 models improved over InSilicoVA/InterVA-5 the most for chronic respiratory diseases (0.74-0.94 PCCC difference), and the least for Malaria (0.09-0.17 PCCC difference). All models had >0.7 PCCC for maternal conditions (0.79-0.99 PCCC), while unspecified infections, malaria, and ill-defined CODs models with <0.5 PCCC. GPT-4 had performance improvements >0.2 PCCC compared to all other models for cancers (+0.25-0.36 PCCC), stroke (+0.27-0.45 PCCC), and diarrhoeal diseases (+0.37-0.51 PCCC), while

GPT-3.5 had similar improvements for liver and alcohol related diseases (+0.27-0.52 PCCC).

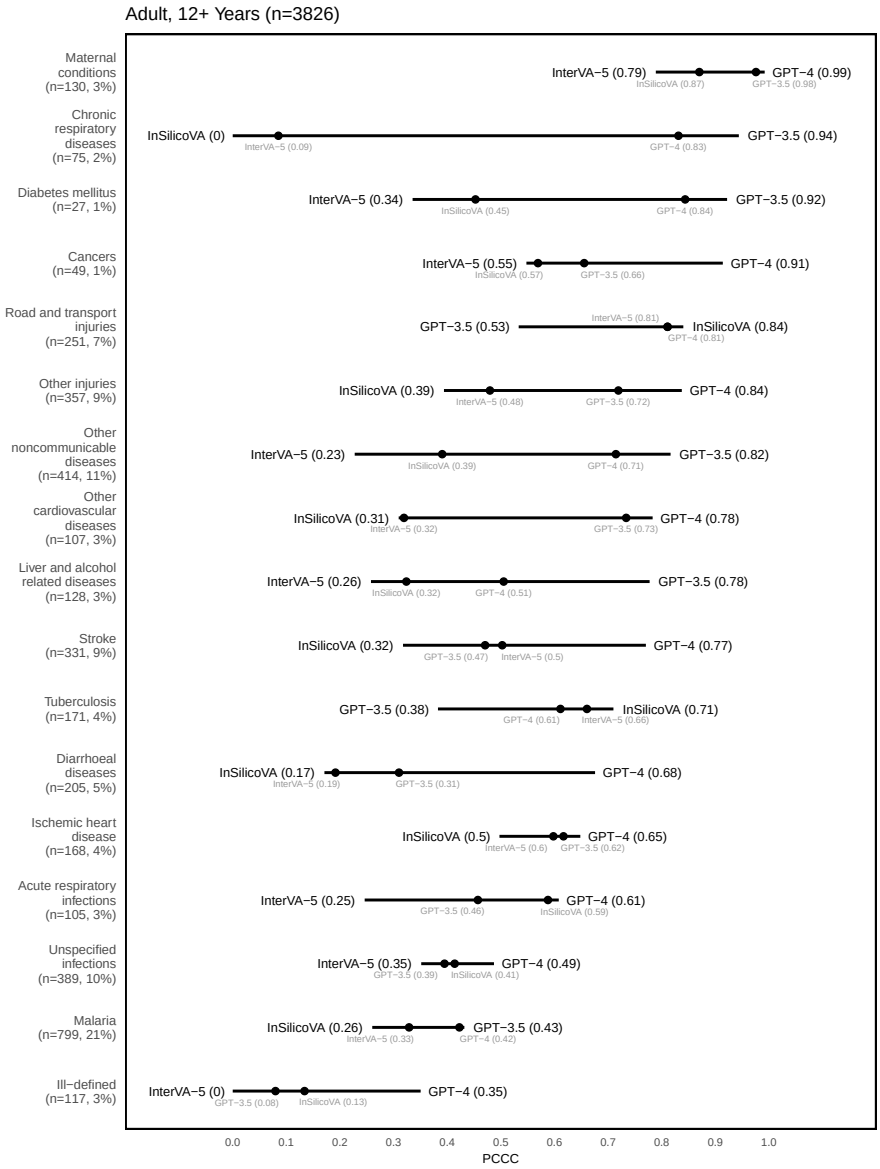

Fig. 4 Model performance for adult records by COD.

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

Figure **??** presents individual performances for each of the models by 8 child CODs, excluding congenital anomalies due to low sample size (n=1, <1%). GPT-4 had the highest individual performance for 4 of 8 CODs (0.65-0.94 PCCC), GPT-3.5 for 3 CODs (0.44-0.88 PCCC), and InSilicoVA for 1 COD (0.78 PCCC). InterVA-5 had the lowest performance for 4 of 8 CODs (0.09-0.79 PCCC), InSilicoVA for 3 CODs (0-0.35 PCCC), and GPT-3.5 for 1 COD (0.58 PCCC).

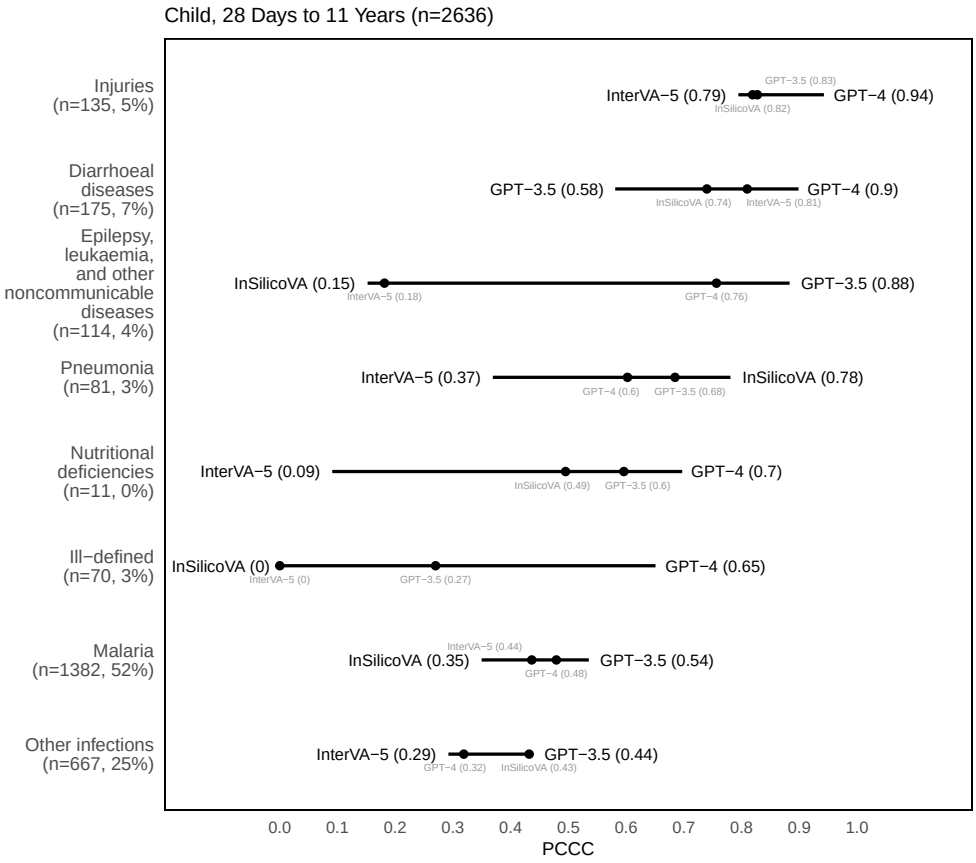

Fig. 5 Model performance for child records by COD.

3.4 Performance for 477 Neonatal Records (Under 28 Days)

Across 5 neonatal CODs (excluding congenital anomalies and other due to small sizes of 2 (<1%) and 5 (1%) respectively), the highest PCCC was observed for GPT-4 for 3 of the 5 CODs (GPT-3.5 identical for prematurity and low birthweight), GPT-3.5 for 2 of the 5 CODs (InSilicoVA identical for stillbirth), and InSilicoVA for neonatal infections (Figure **??**). Stillbirth had small differences between minimum and maximum PCCC at 0.07. The other 4 CODs had large differences between minimum and maximum PCCC ranging from 0.46 to 0.64. Each neonatal COD had between 23 (5%) 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 ill-defined.

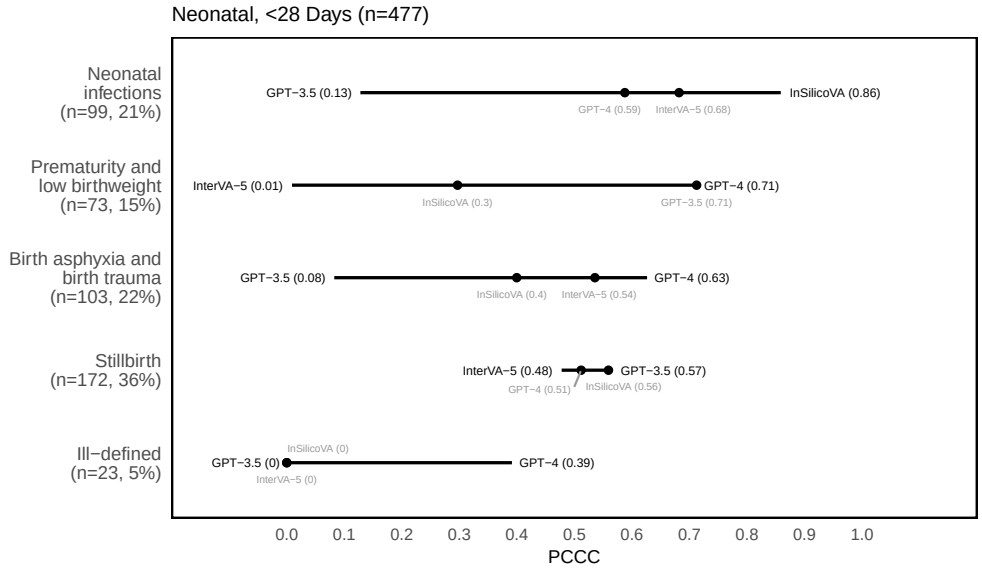

Fig. 6 Model performance for neonatal records by COD.

4 Discussion

This section discusses and summarizes the results from Section **??**. Advantages and disadvantages of using GPT-3.5, GPT-4, InterVA-5, and InSilicoVA models for assigning CODs are discussed in Sections **??** and **??**. Limitations of the study are mentioned in Section **??**, while opportunities and future work are detailed in Section **??**.

4.1 Advantages

This section identifies advantages of performant models for assigning CODs. Section **??** details the strategic use of models for particular CODs. Section **??** details the resource efficiency of GPT and InSilicoVA models for assisting in physician COD assignment. Section **??** notes the strength of using natural language text in GPT models compared to structured data (e.g. questionnaires) for physician COD assignment.

4.1.1 Performance for Particular Causes of Death

At the population level, overall performances for all causes were similar ($0.7 \leq$ CSMF ≤ 0.79), but not close ($\geq$0.8 CSMF), to physicians at 0.74-0.79 CSMF. At the individual level, performance by PCCC varies across CODs, and for most CODs, GPT models (GPT-4 and GPT-3) performed better than InSilicoVA and InterVA-5 (Table **??**). However, there were CODs where InSilicoVA performed better. For example, neonatal infections at 0.87 PCCC versus 0.23 for GPT-3.5, and adult tuberculosis at 0.72 versus 0.39 for GPT-3.5. For CODs with high performance ($\geq$0.8 PCCC) (Table **??**), the results suggest that GPT models (and InSilicoVA for particular CODs) may assign CODs that are close physicians, when physicians would agree on the COD. Thus, when applying models to assign CODs, it may help to target CODs with a combination of models that performed well or close to physician assignment.

Across age ranges, GPT and InSilicoVA models had performances below 0.8 PCCC, where ages 10-24 years and 7-27 days had the highest performance ($\geq$0.7 PCCC). For

adult age ranges, performance generally decreased as age increased, which suggested that models had difficult assigning CODs for older than younger adults (Figure **??**). For child and neonatal age ranges, the performance improves as the age increases, suggesting less difficulty in COD assignment when children and neonates are more developed (Figures **??** and **??**). As the models did not perform well for any particular age range, it may not be suggested to apply specific models to target cases by age range. However, the general patterns (increases and decreases of performance in relation to age) are useful in comparing expected patterns in relation to physician assignment.

Table 1 Models with performances close to physician COD assignment ($\geq$0.8 PCCC) across age groups

Cause of Death	Cases	Best Model	PCCC
Adult (n=3826, 1303 (34%) $\geq$ 0.8 PCCC)			
Maternal conditions	130 (3%)	GPT-4	0.99
Chronic respiratory diseases	75 (2%)	GPT-3.5	0.94
Diabetes mellitus	27 (1%)	GPT-3.5	0.92
Cancers	49 (1%)	GPT-4	0.92
Road and transport injuries	251 (7%)	InSilicoVA	0.84
Other injuries	357 (9%)	GPT-4	0.84
Other noncommunicable diseases	414 (11%)	GPT-3.5	0.82
Child (n=2636, 505 (19%) $\geq$ 0.8 PCCC)			
Injuries	135 (5%)	GPT-4	0.95
Diarrhoeal diseases	175 (7%)	GPT-4	0.91
Epilepsy, leukaemia, and other noncommunicable diseases	114 (4%)	GPT-3.5	0.89
Pneumonia	81 (3%)	InSilicoVA	0.8
Neonatal (n=477, 99 (21%) $\geq$ 0.8 PCCC)			
Neonatal infections	99 (21%)	InSilicoVA	0.87

4.1.2 Highly Scalable and Available

The models in this study can assist physicians in assigning CODs in a variety of ways due to low costs and speed of COD assignment. Similar to differential diagnoses, GPT and InSilicoVA models can offer more alternative COD assignments for physicians to consider, [**?**], which can potentially help lower the number of records with ill-defined

causes or reduce disagreement in physicians. At the time of this study, running GPT-3.5 cost ∼$1.6 USD ($0.5 per one million tokens), GPT-4 cost ∼$115 USD ($30 per one million tokens), and InSilicoVA had no costs on 6939 physician agreed records [**?**]. These costs are likely lower than physicians (e.g. less than $3 USD per house in India [**? ?**]), while over 10,000 records can be coded in under a day. When physicians are unavailable, GPT and InSilicoVA models can be a cost-efficient alternative to code large amounts of records for population estimates of CODs. However, care needs to be taken to apply these models only for certain CODs where models perform well, such as in Table **??**. In addition, these models can also help divert physician resources to cases that are more difficult to code or require more attention For example, physicians can validate cases, such as road traffic injury where models performed well at 0.84 PCCC, while spending more time on cases that related to acute respiratory infections, where the models performed poorly at 0.62 PCCC (Figure **??**). Lastly, as models can assign CODs on-demand, there is potential for models to provide CODs during the data collection process, and for GPT models to guide the surveyor to ask additional questions for improving narrative quality.

4.1.3 Natural Language Flexibility

All models did not require training data to assign CODs, which allowed them to be used without domain expertise and supplying training datasets. The main advantage to GPT-3.5 and GPT-4 was the use of natural language text as input and output. Compared to InterVA-5 and InSilicoVA, GPT models were able to assign COD codes in ICD-10, as physicians do, and potentially assign CODs in more broad categories depending on the prompts. In comparison, InterVA-5 and InSilicoVA relied on structured input and output data from WHO VA 2016 questionnaires, and assigned CODs in WHO VA 2016 codes only. This required that these codes and forms be maintained with conversions between different form (e.g. WHO VA 2012 to WHO VA

2016) and code standards (e.g. WHO VA 2016 to ICD-10), which reduces interoperability and comparability when other models and VA systems do not use the same or interchangeable standards. Thus, GPT models did not require strict formats for training and testing data, which can capture latent and more ambiguous patterns (e.g. health-seeking behaviours and social issues) outside the scope of WHO VA codes and forms [**?** **?**]. For example, GPT-3.5 and GPT-4 had higher performance (+0.14-0.67 PCCC) than InterVA-5 and InSilicoVA for more ambiguous CODs (e.g. other/unspecified infections, ill-defined, other cardiovascular diseases). GPT models also performed better (+0.57 PCCC) on CODs with a small number of cases, such as nutritional deficiencies (n=11) and diabetes mellitus (n=27), which may not have enough cases for questionnaires to capture statistical patterns, but may possibly have richer contextual information from articles, web sources, or books that GPT models can leverage.

4.2 Disadvantages

This section discusses the disadvantages of GPT models for COD assignment. Section **??** identifies issues in reproducing GPT outputs for repeated runs on the same records and lack of up-to-date information, while Section **??** discusses the resource intensive infrastructure required by GPT and its relation to data privacy.

4.2.1 Reproducibility and Timeliness

The GPT models in this study had the temperature parameter set to 0 for more reproducible results. A short experiment in Appendix **??** revealed that GPT-3.5 assigns the same COD for the same record only more than 60% of the time, based on repeated runs on a sample of 100 records. Although more extensive testing is needed, this suggests that GPT models do not always assign the COD for the same case on multiple runs, which may pose issues in reproducibility and reliability. For example, GPT models may achieve correct COD assignments solely due to random chance, but are difficult to test with large numbers (e.g. 10,000) of reruns due to costs (e.g. costs increased 10

fold per record when rerun 10 times). In comparison, InterVA-5 and InSilicoVA are open source and free, and assign deterministic CODs with probabilities for each alternative COD, which offers more reproducible and reliable COD assignments despite lower performance overall. In addition, all models were trained on historical data up to particular points in time, which may not utilize the most up-to-date data available (e.g. latest online articles, social media, or books for GPT models).

4.2.2 Infrastructure and Data Privacy

GPT-3.5 and GPT-4 models required large computing infrastructure to train and run, which was not possible to run on local computers, or setup due to costs. This poses issues with data privacy as sensitive data (e.g. identifying information) needs to be sent to company servers, which can be collected by companies (e.g. OpenAI) and misused [**?**]. For example, GPT models use prompts, which contain the narrative data, to assign CODs. The data in the promots may be unknowingly collected and misused by companies (e.g. companies) or their users (e.g. malicious prompts) [**? ?**]. In contrast, InterVA-5 and InSilicoVA can be run on local computers, which allows data to stay with the owner to protect data privacy, without reliance on external services.

4.3 Limitations

This section identifies limitations in this research in the context of GPT models. Section **??** identifies the omission of ICD-10 performance evaluations. Section **??** discusses data limitations, such as small samples, exclusion of disagreements, and exploration of incorrectly assigned cases. Section **??** mentions the need for parameter tuning and evaluation of consistency and multiple COD assignments.

4.3.1 ICD-10 Performance Evaluation

For the scope of this study, all models were evaluated for their performance in broad CGHR-10 COD categories as opposed to more specific ICD-10 codes. However, in

practical cases, physicians assign more specific ICD-10 codes rather than broader COD categories. InterVA-5 and InSilicoVA assigned broader WHO VA codes, and were unable to assign ICD-10 codes, as the number of cases for specific ICD-10 codes may not have enough cases for training statistical models. GPT models were able to assign ICD-10 codes, but may possibly result in lower performance as even physicians do not agree completely on ICD-10 codes, and broader categories (CMEA-10 codes in Additional file 2) were used to assign equivalency or agreement.

4.3.2 Small Samples, Disagreements, and Misclassifications

Performance evaluations and analyses were omitted for CODs with less than 10 cases (e.g. congenital anomalies, suicide), in which models may perform well on, but have limited evidence to make conclusive insights on performance. In this study, only records where physicians agreed on the COD assignment were included in performance evaluations. However, the performance of GPT-4 only dropped slightly from 0.61 (agreed) to 0.53 (disagreed) PCCC while records almost doubled, which suggests that GPT models may potentially be able to assign agreeable CODs without reconciliation or adjudication for cases where physicians disagreed (Figure **??**). In relation, GPT may be used to explore prompts that may possibly reconcile or adjudicate records in disagreement. Finally, an exploration of misclassifications was not conducted for this study, but may yield useful insights. For example, a non-comprehensive exploration was conducted in Appendix **??** on misclassified GPT-4 records for neonatal infections, which found potential issues with the categorization of CGHR-10 codes, order of information in narratives, and guidelines of COD assignments. More records may also provide more conclusive evidence of model performance results as 6939 records may be inadequate evidence at the country level compared to more comprehensive studies (e.g. the million death study [**?**]).

4.3.3 Model Tuning, Consistency, and Multiple Outputs

GPT-3.5 and GPT-4 models used default parameters with the exception of setting the temperature to 0 for more consistent results. However, the temperature and other potential settings may be adjusted to possibly improve performance [**?**]. In addition, GPT models may possibly produce inconsistent results even with the temperature set to 0 as discussed in Section **??**. Thus, it is important to also test the reliability and consistency of GPT outputs to avoid coincidental results due to randomness [**? ? ?**]. InterVA-5 and InSilicoVA were able to provide multiple COD assignments with probabilities for each. GPT models can be prompted to produce more than one COD assignment, but was not explored in this study. This may be useful to evaluate the performance of suggested alternative COD assignments, which can help physicians reach agreement or reduce ill-defined assignments.

4.4 Opportunities

This section discusses research opportunities to improve GPT models for assigning CODs. Section **??** discusses the potential to improve GPT models with prompt engineering and exploration of misclassified records. Section **??** identifies an opportunity to integrate GPT, InterVA-5, and InSilicoVA models into VA systems for improving physician COD assignment.

4.4.1 Prompt Engineering and Custom Models

Prompt engineering, the act of designing prompts to guide GPT models for better results [**?**], presents an important research opportunity that may improve performance of GPT models for COD assignment. As mentioned in Section **??** and exemplified in Appendix **??**, an analysis of misclassified records may yield insights on adjusting prompts to assign more correct CODs in relation to physicians. In addition, subsequent prompts and examples can be used to add in correctional instructions and refine

results, while additional information from the questionnaire and physician VA manuals may add better context [**?**]. Sensitivity analyses may be conducted to assess the effects on performance and consistency of results from modified prompts on a COD basis. GPT models may also be customized to specific domains or contexts, where objectives, behaviours, extra data, privacy, and evaluation tests can be adjusted to produce custom models that perform better in targeted domains or circumstances [**?**].

4.4.2 Computer Assisted Verbal Autopsy

Another research opportunity is in the integration of GPT, InterVA-5, and InSilicoVA models into VA systems to assist physicians in COD assignment. In dual-coded VA systems (described in Section **??**), two physicians are randomly assigned to each record and require second inspections of each other's assignment (reconciliation) and evaluation by a third more senior physician if their assignments do not agree. As mentioned in Section **??**, suggestion of alternative assignments from GPT and InSilicoVA models potentially reduces the disagreement between physicians, and ill-defined records, while allowing physicians to focus on more difficult records. Thus, model suggestions can be integrated into VA systems by presenting COD suggestions to physicians after their initial COD assignment, which allows them to consider alternative assignments and possibly revise their assignments based on the suggestions. At step 2 in Figure **??**, GPT, InterVA-5, and InSilicoVA models can suggest COD assignments to consider, providing the option in step 2b to revise or proceed with their initial assignment. Our future work will be a first step in computer assisted verbal autopsy, assessing the effects of these model suggestions on improve VA data quality (e.g. increase in agreed records, reduction of ill-defined deaths). In preparation, we have integrated GPT-4, InterVA-5, and InSilicoVA model suggestions into our HEAL-SL study after survey round 2 [**?**].

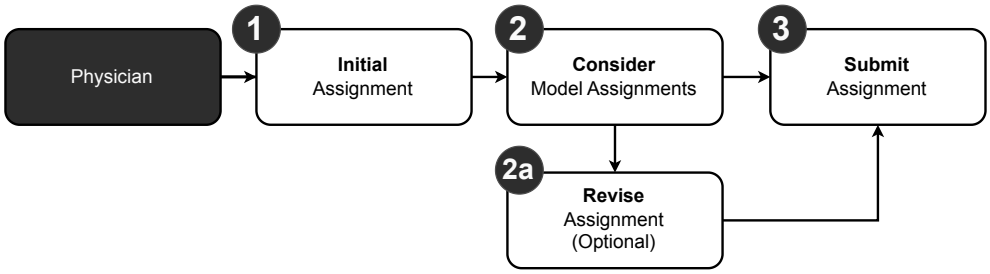

Fig. 7 Model suggestions integrated in the physician assignment process.

5 Conclusion

This study evaluates the performance of GPT-3.5, GPT-4, InterVA-5, and InSilicoVA models compared to physicians for assigning CODs for 6939 VA records in Sierra Leone (2019-2022). At the population level, all models had CSMF accuracies of 0.7-0.79. At the individual level, GPT-4 had the best performance (0.61 PCCC), followed by GPT-3.5 (0.58 PCCC), InSilicoVA (0.44 PCCC), and InterVA-5 (0.45 PCCC). When evaluating performance by COD, 7 of 19, 4 of 10, and 1 of 7 adult (12-69 years), child (28 days to 11 years), and neonatal (under 28 days) CODs respectively had models with $\geq$0.8 PCCC (see Table **??**). Performance decreased from 0.72 to 0.59 PCCC as adults aged, and increased from 0.5 to 0.7 PCCC as children and neonates developed. Thus, GPT and InSilicoVA models were comparable to physicians for particular CODs (e.g. maternal conditions, injuries, diarrhoeal diseases, pneumonia), but not across age ranges. Advantages of GPT (and InSilicoVA for some CODs) models include being highly scalable and available, which allows the suggestion of multiple alternative COD assignments, and reduction of time spent on non-complex cases to assist physicians in COD assignment. In addition, GPT models provide flexible natural language input and output, capturing latent patterns (e.g. health-seeking and social issues) that potentially lead to their overall high performance compared to InterVA-5 and InSilicoVA. However, GPT models do not always assign the same COD for the same record on multiple runs, are trained on data from a specific time period, and

require large computing infrastructure, leading to disadvantages in reliability of COD assignments, timeliness of up-to-date information, and data privacy issues. Limitations of this study were small sample sizes in relation to rarer CODs and other national level studies, need for more comprehensive exploration of disagreed and misclassified records, and lack of experimentation of model parameters, output consistency, and multiple COD assignments. Research opportunities include refining GPT models using prompt engineering and custom models, and future work towards computer assisted verbal autopsy, where GPT and other models are used to assist physician COD assignment by offering multiple alternative assignments, improving physician agreement on COD assignment and reducing ill-defined deaths. GPT-4, InterVA-5, and InSilicoVA has been integrated into future rounds of the HEAl-SL study to assist physicians with alternative COD assignments. Future work in evaluating the effectiveness of these models to reduce disagreements among physicians and ill-defined deaths will be a step forward in more accurate and efficient VA systems.

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

TBD.

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) and at https://github.com/cghr-toronto/healsl-gpt-paper.

Verbal Autopsy (VA) data by age group and survey rounds 1 and 2:

- Adult VA (Round 1):
  https://openmortality.org/data/healsl_rd1_adult_v1
- Adult VA (Round 2):
  https://openmortality.org/data/healsl_rd2_adult_v1
- Child VA (Round 1):
  https://openmortality.org/data/healsl_rd1_child_v1
- Child VA (Round 2):
  https://openmortality.org/data/healsl_rd2_child_v1
- Neonatal VA (Round 1):
  https://openmortality.org/data/healsl_rd1_neo_v1
- Neonatal VA (Round 2):
  https://openmortality.org/data/healsl_rd2_neo_v1

Narrative data by age group and survey rounds 1 and 2:

- Adult Narratives (Round 1):
  https://openmortality.org/data/healsl_rd1_adult_narrative_v1
- Adult Narratives (Round 2):
  https://openmortality.org/data/healsl_rd2_adult_narrative_v1
- Child Narratives (Round 1):
  https://openmortality.org/data/healsl_rd1_child_narrative_v1
- Child Narratives (Round 2):
  https://openmortality.org/data/healsl_rd2_child_narrative_v1
- Neonatal Narratives (Round 1):
  https://openmortality.org/data/healsl_rd1_neo_narrative_v1
- Neonatal Narratives (Round 2):
  https://openmortality.org/data/healsl_rd2_neo_narrative_v1

Cause of death code mappings to convert between ICD-10, WVA-2016, and CGHR-10 codes:

- ICD-10 to CGHR-10:
  https://openmortality.org/data/cghr10_v1
- WVA-2016 to ICD-10:
  https://openmortality.org/data/icd10_wva2016_v1

Result files:

- Model and physician COD assignments:
  https://github.com/cghr-toronto/healsl-gpt-paper/blob/main/data/healsl_rd1to2_cod_v1.csv

- Metrics for model COD assignments: https://github.com/cghr-toronto/healsl-gpt-paper/blob/main/data/healsl_rd1to2_metrics_v1.csv
- GPT-3.5 repeated runs results in Appendix **??**: https://github.com/cghr-toronto/healsl-gpt-paper/blob/main/data/healsl_rd1to2_rapid_gpt3_sample100_v2b.csv

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 Data Exploration

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 **??**) 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 **??**). 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 **??**). 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 **??**). 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 **??**). 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 **??**). 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 **??**).

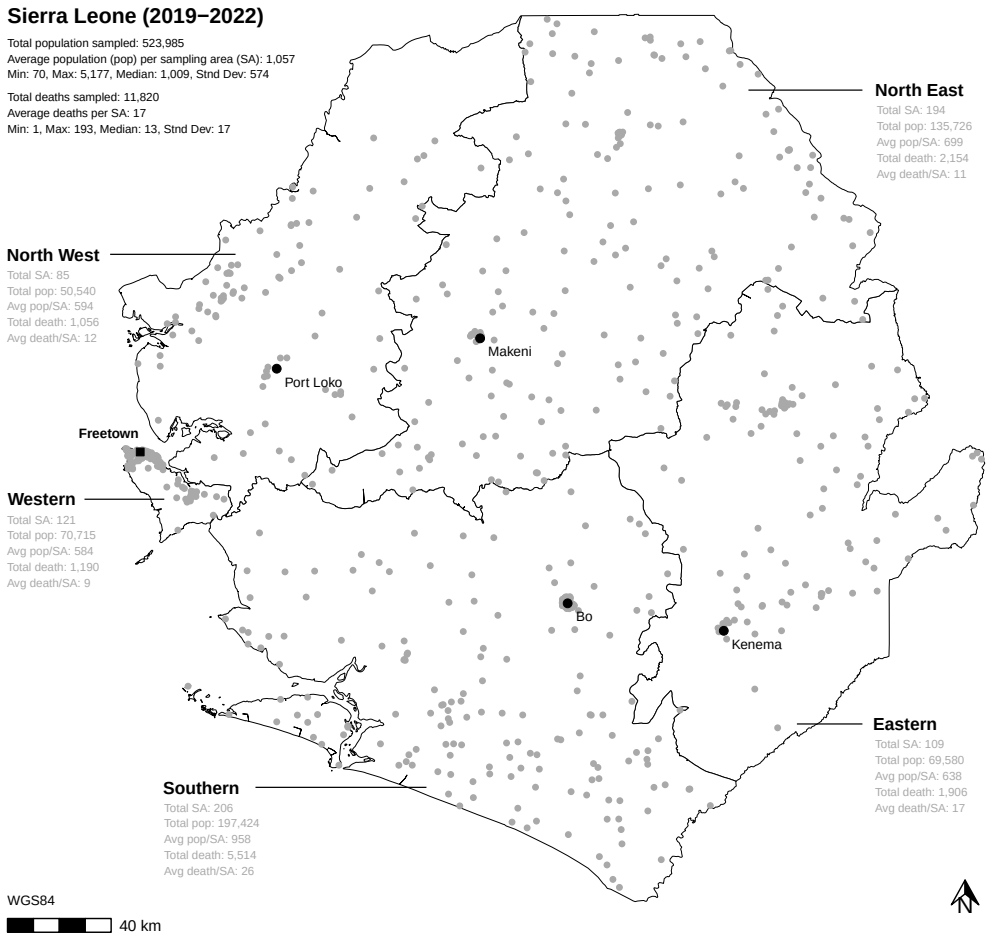

Fig. A1 Case study data sampling areas.

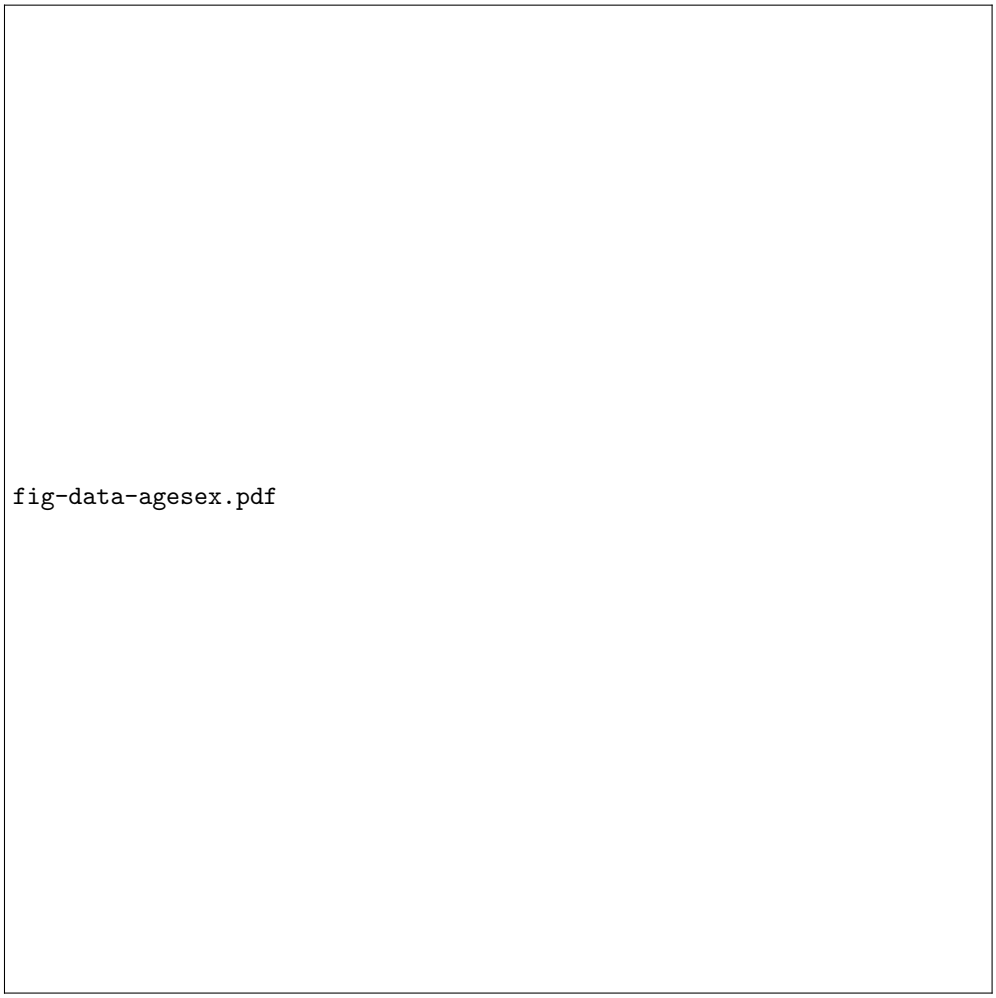

Fig. A2 All records by age group and five-year age range.

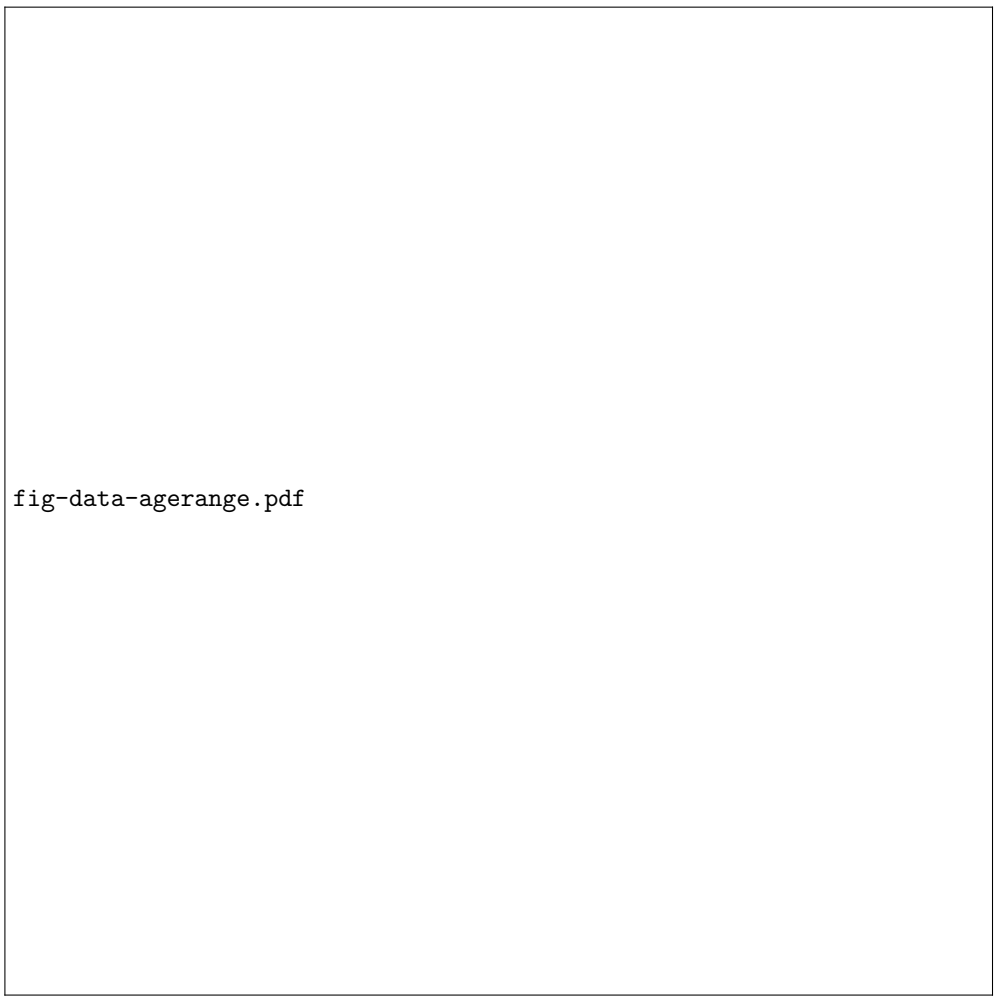

Fig. A3 All records by five-year age range.

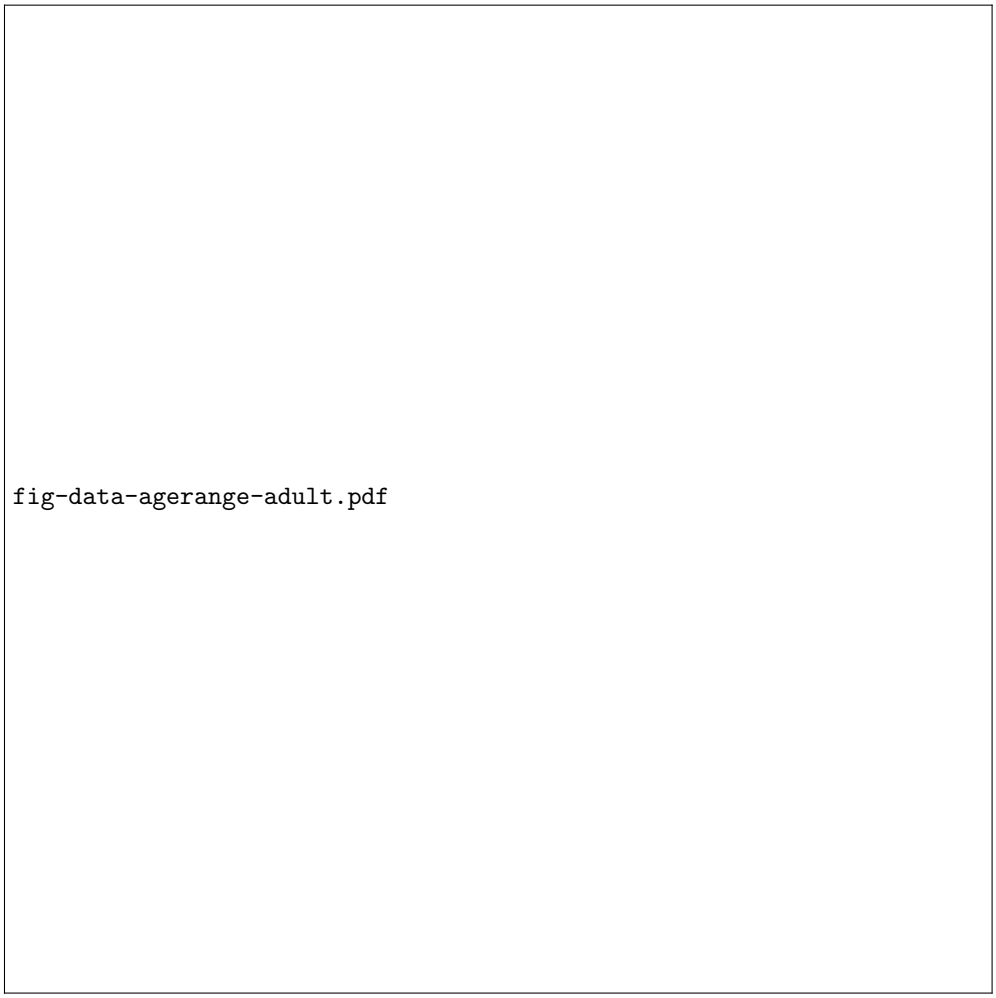

Fig. A4 Adult records by five-year age range.

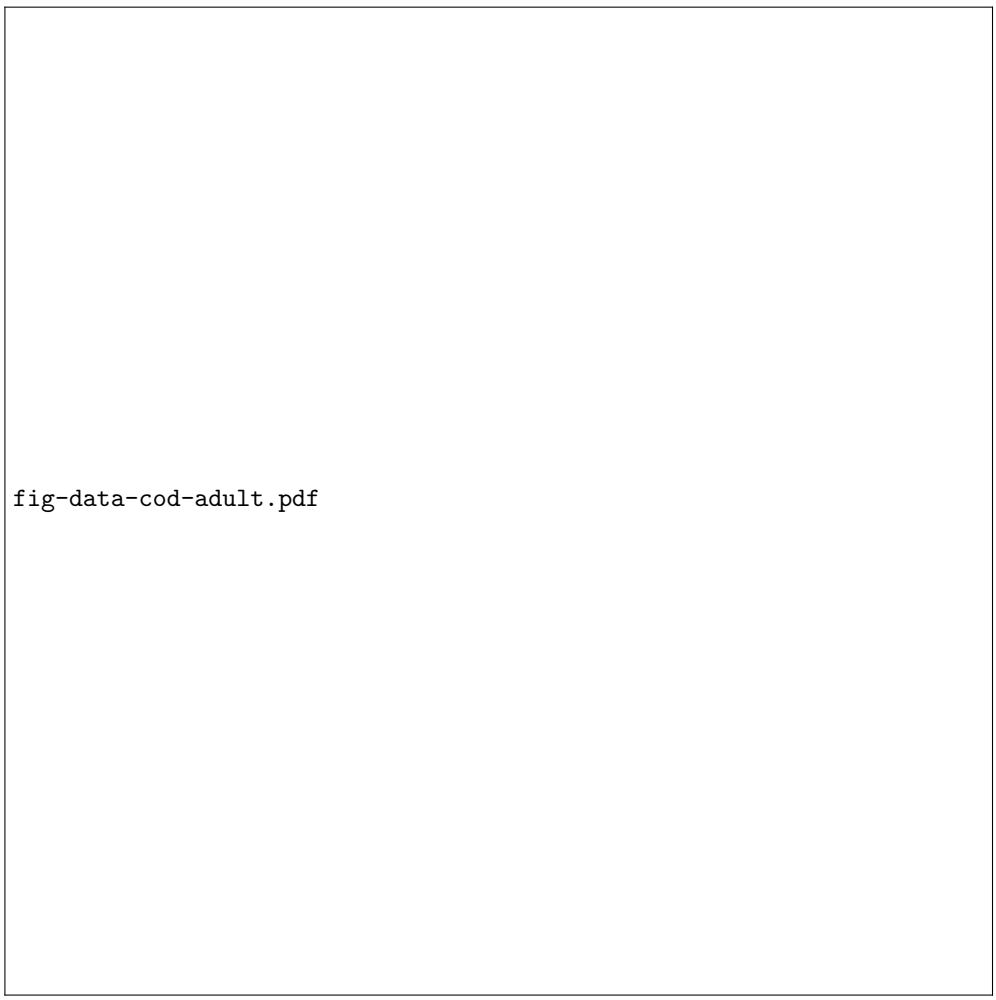

Fig. A5 Adult records by COD.

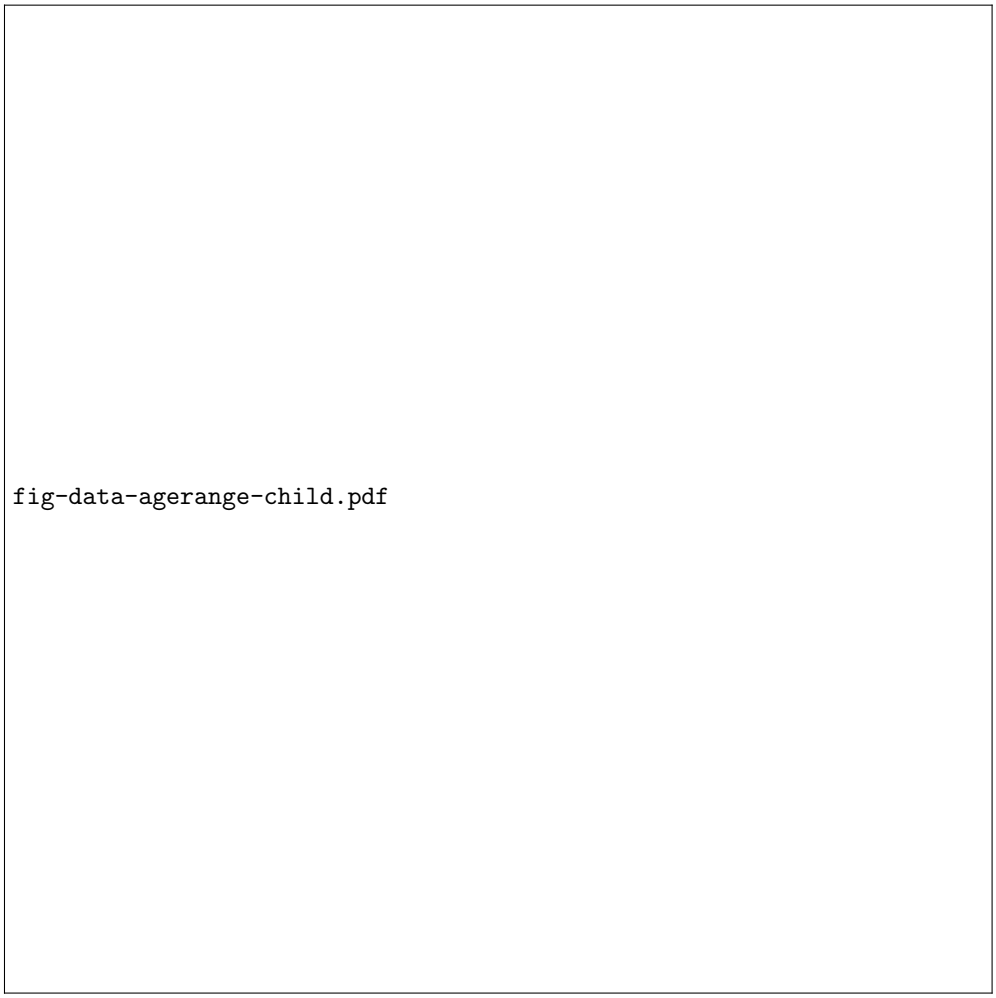

Fig. A6 Child records by five-year age range.

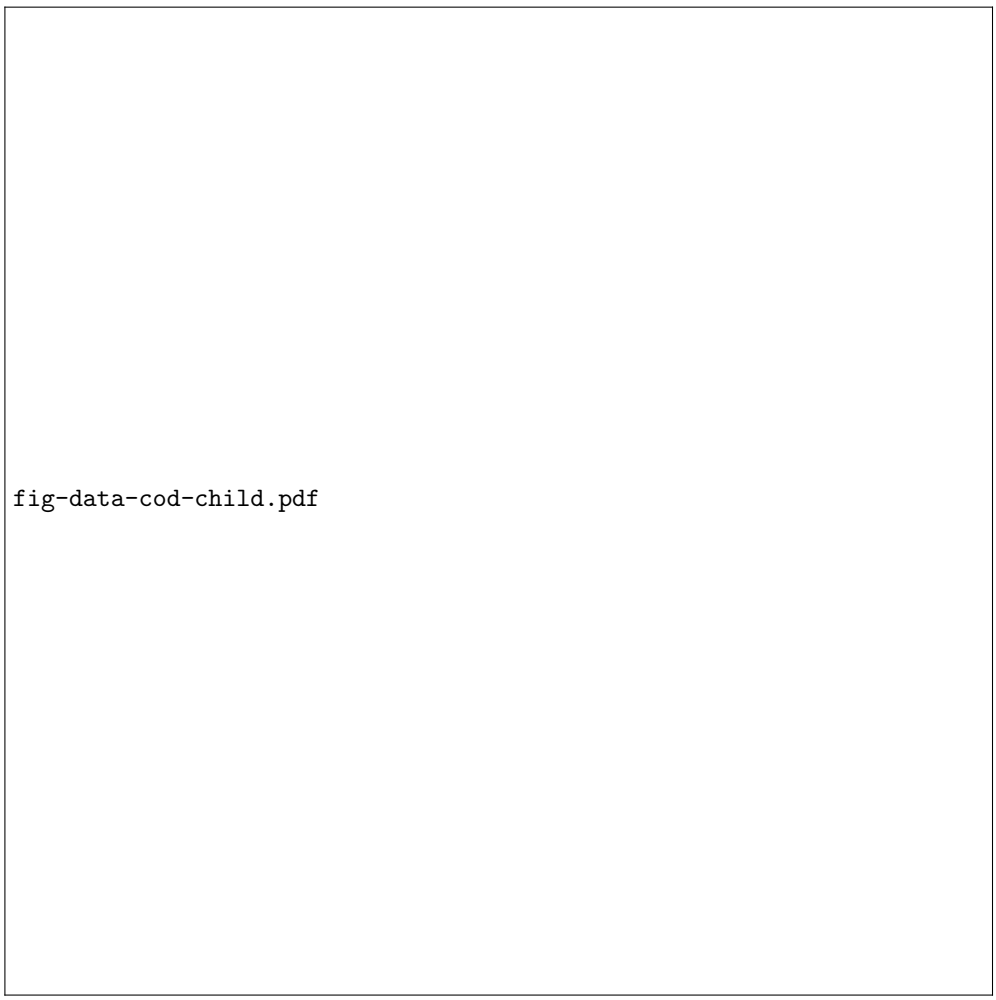

Fig. A7 Child records by COD.

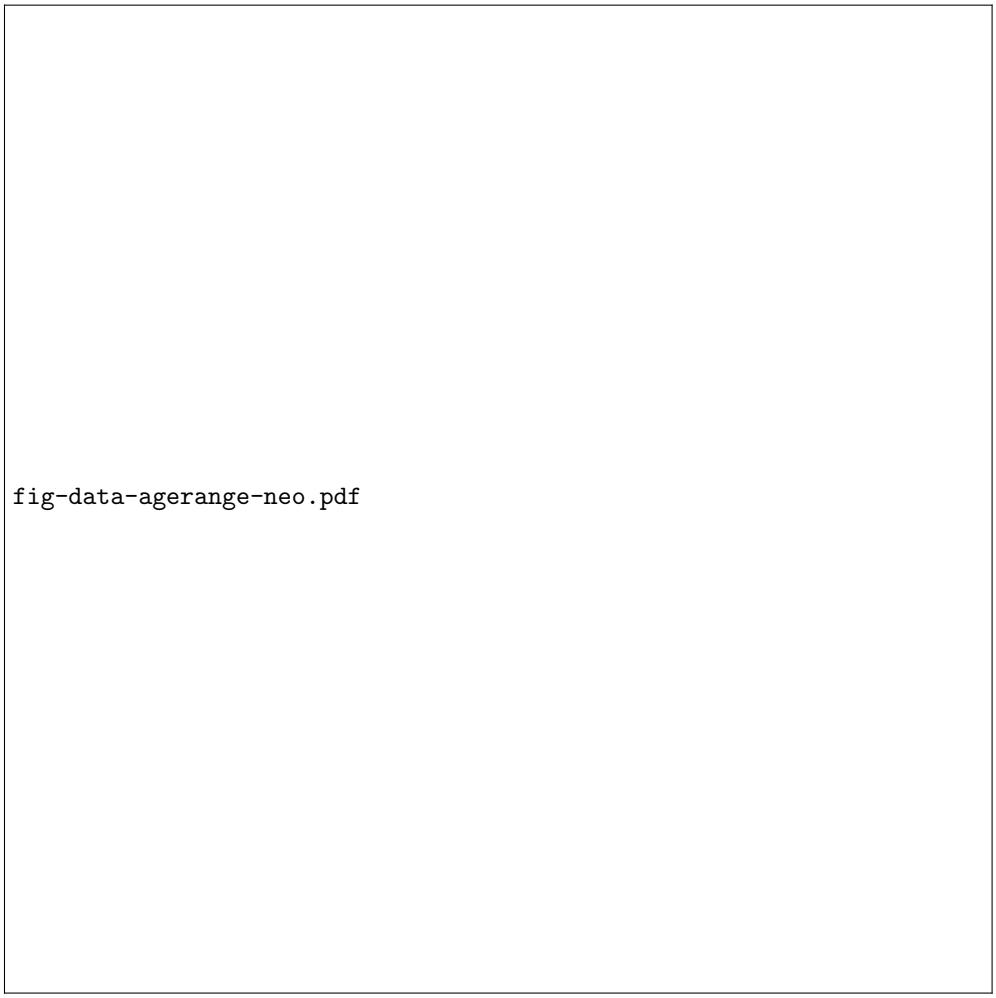

Fig. A8 Neonatal records by five-year age range.

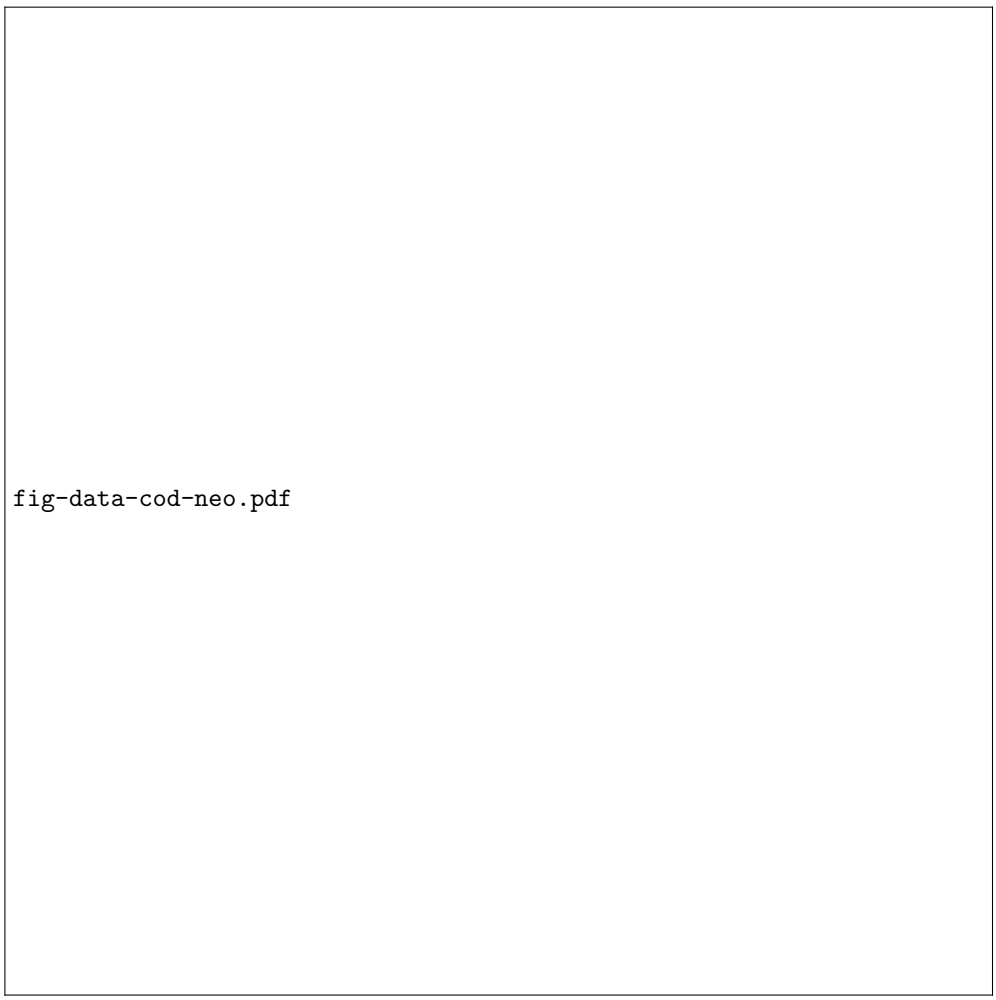

Fig. A9 Neonatal records by COD.

Appendix B Details on Methods

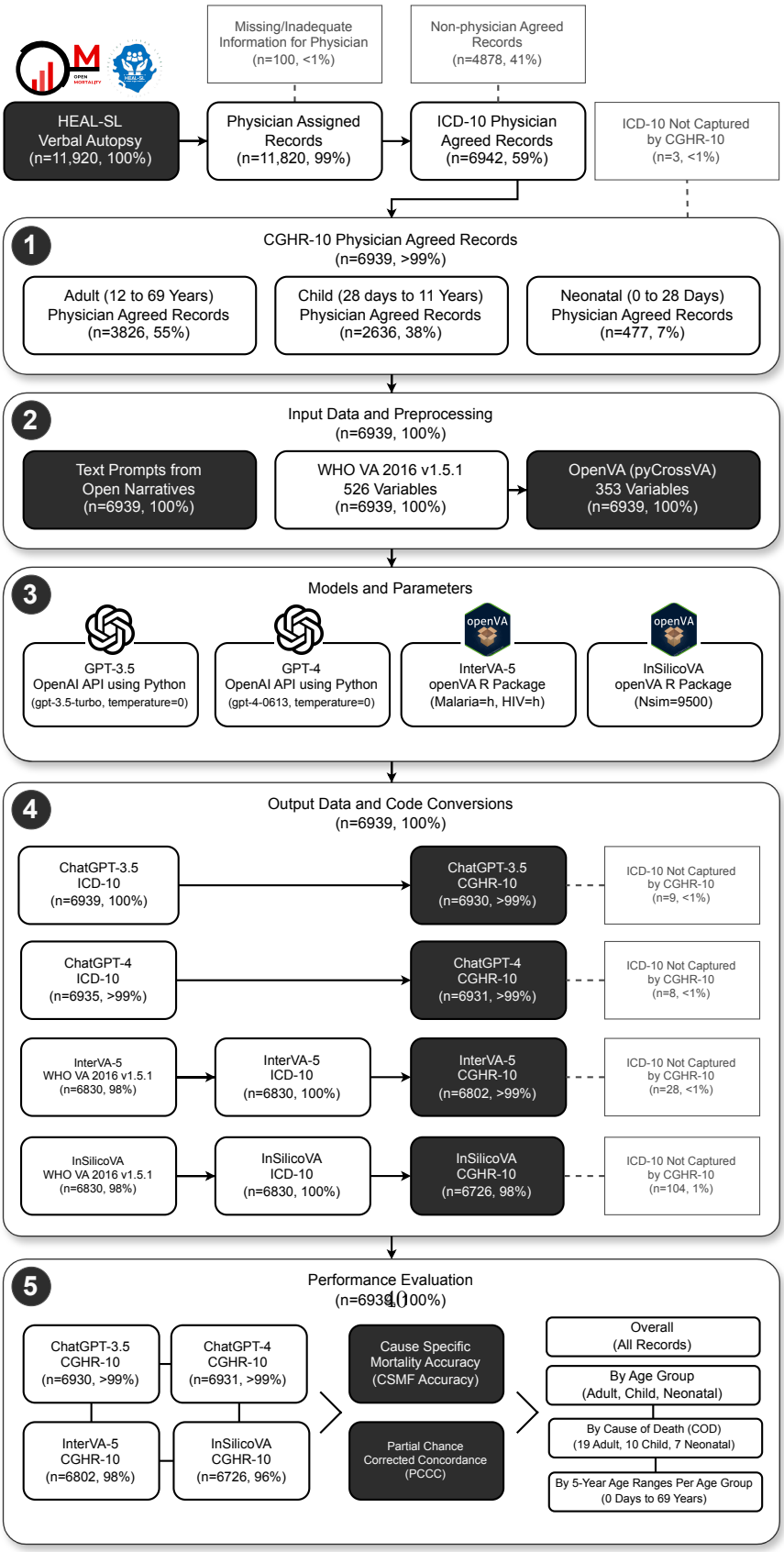

Fig. B10 Case study methods.

B.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 [**?**]. 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 **??**. Then, the $CSMFMaximumError$, representing the worst possible model, is calculated using Equation **??**. Finally, the CSMF accuracy is given by Equation **??**, 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.

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

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

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

B.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 [**?**]. PCCC is given by Equation **??**, 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 **??**. 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.

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

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

Appendix C Experiment on Repeated Runs of GPT-3.5

A short experiment was conducted to test the consistency of GPT-3.5 outputs repeated on the same record. 100 records, sampled randomly with approximately equal proportions across age groups, CODs, and survey rounds 1 and 2, were used to test repeated runs of GPT-3.5. Each record from the 100 records was rerun 10 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 **??**. 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 C1 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 D 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 **??**, 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 D2 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%)