

001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046

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

Richard Wen^{1*}, Anteneh Tesfaye Assalif^{1,2}, Andy Sze-Heng Lee¹,
Rajeev Kamadol¹, Asha Behdinan¹, Ronald Carshon-Marsh¹,
Thomas Kai Sze Ng¹, Patrick Brown¹, Prabhat Jha¹,
Rashid Ansumana²

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.

²School 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;
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 representative Causes of Death (CODs). Current computer models for COD assignment in VAs perform similar to physicians at the population level, but poorly at the individual level, due to focuses on questionnaire data and neglecting free text from narratives. Recently, a

047 large language model called ChatGPT-4 demonstrated human-level performance
048 on professional and academic exams. ChatGPT-4 shows promise in assigning
049 CODs similar to physicians, but has yet been examined for assigning CODs using
050 VA narratives.

051 **Methods:** 6939 VA records in Sierra Leone from 2019 to 2022 were used to
052 compare four computer models, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA, to
053 physician COD assignment at population and individual levels. Narratives were
054 used for GPT-3.5/4, while questionnaires were used for InterVA-5/InSilicoVA.
055 COD assignments were grouped into general COD categories consisting of 19, 10,
056 and 7 categories for adult, child, and neonatal age groups. Cause Specific Mor-
057 tality Fraction (CSMF) accuracy and Partial Corrected Concordance (PCCC)
058 were used to compare models to physicians at population and individual levels
059 respectively. CSMF and PCCC were evaluated overall and by COD, age group,
060 and age ranges.

061 **Results:** GPT-4 had the best performance overall (0.61 PCCC), followed by
062 GPT-3.5 (0.56 PCCC), and InSilicoVA/InterVA-5 (0.44 PCCC). GPT-4 had
063 the best performance for adult and neonatal records (0.64 and 0.58 PCCC),
064 with GPT-3.5 for child records (0.54 PCCC). All models' performances trended
065 upwards from 1 month to 14 years (\sim 0.1-0.75 PCCC) and downwards from 15-69
066 years (\sim 0.7-0.35) of age. GPT-4, GPT-3.5, and InSilicoVA had the highest per-
067 formances for 17, 9, and 4 of all 30 CODs respectively. At the population level,
068 all models had CSMF accuracies between 0.74-0.79.

069 **Conclusion:** All models performed well at the population level, while GPT-
070 3.5/4 and InSilicoVA performed well at the individual level for some CODs.
071 GPT models have yet to replace physician coding, but made improvements over
072 InSilicoVA and InterVA-5. Our research lays the foundation for future work in
073 computer assisted VA, where physicians utilize alternative COD assignments from
074 computer models to help reduce ill-defined codes and physician disagreement.

075 **Keywords:** Cause of Death, Physician Coding, Verbal Autopsy, GPT, AI, LLM

076

077

078 1 Background

079

080 In 2019, 41 million people die prematurely from noncommunicable diseases every year,
081 accounting for 74% of all deaths globally [1]. Most of these deaths are preventable,
082 but require adequate resource allocation, guided by evidence, to implement effective
083 interventions and policies that target populations at risk [2]. Thus, reliable counts and
084 diagnoses of deaths enable decision makers to identify populations at risk to save lives
085 and reduce premature deaths worldwide [3–6]. However, most low-income countries
086 do not have data on deaths or have registered less than half of the deaths in their
087

088

089

090

091

092

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].	093 094 095 096 097 098 099 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138
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. This component has been criticized to be difficult to reproduce due to reliance on physician assignment [15–19]. As an alternative to physician assignment, computer models, such as InterVA [20] and InSilicoVA [17], have 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 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 [26–28].	124 125 126 127 128 129 130 131 132 133 134 135 136 137 138
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]. 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	124 125 126 127 128 129 130 131 132 133 134 135 136 137 138

139 GPT-3, which had notable differences in training data sizes of 5 gigabytes to 45 ter-
140 abytes from web sources that resulted in 117 million to 175 billion parameter models
141 [33]. In March 2023, ChatGPT-4 was released with human-level performance on vari-
142 ous professional and academic exams and benchmarks that outperformed ChatGPT-3
143 [34]. Given the limited usage of free-text open narratives in computer models for
144 determining CODs, and recent advances in LLMs that leverage natural language text
145 prompts, we conducted a case study with Sierra Leone deaths from VA in 2019 to 2022
146 to compare four models, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA, to physicians
147 for determining CODs.
148
149

150 2 Methods

151

152 This study details the methods used to compare the COD assignment from four mod-
153 els, GPT-3.5, GPT-4, InterVA-5, and InSilicoVA, to physicians as seen in Figure 1.
154 The initial VA data was filtered for physician agreed records as described in Section
155 2.1. Section 2.2 describes the input and output of the four models for COD assignment,
156 while section 2.3 details the performance evaluation of the models relative to physi-
157 cians using population and individual level metrics. See Appendix A for additional
158 details on the methods used in this study.

159 2.1 Verbal Autopsy (VA) Data

160 Initially, 11,920 records from the HEAL-SL study [35, 36] were collected from dual-
161 coded EVA, where each record was randomly coded by two different physicians that
162 assigned CODs as International Classification of Diseases Revision 10 (ICD-10) codes
163 [37]. For each record, two codes were assigned by two different randomly selected
164 physicians, where codes were evaluated for agreement using Central Medical Evalu-
165 ation Agreement 10 (CMEA-10) codes. CMEA-10 groups a range of similar ICD-10
166 codes together, where if they are in agreement if they are within the same group [38]
167
168

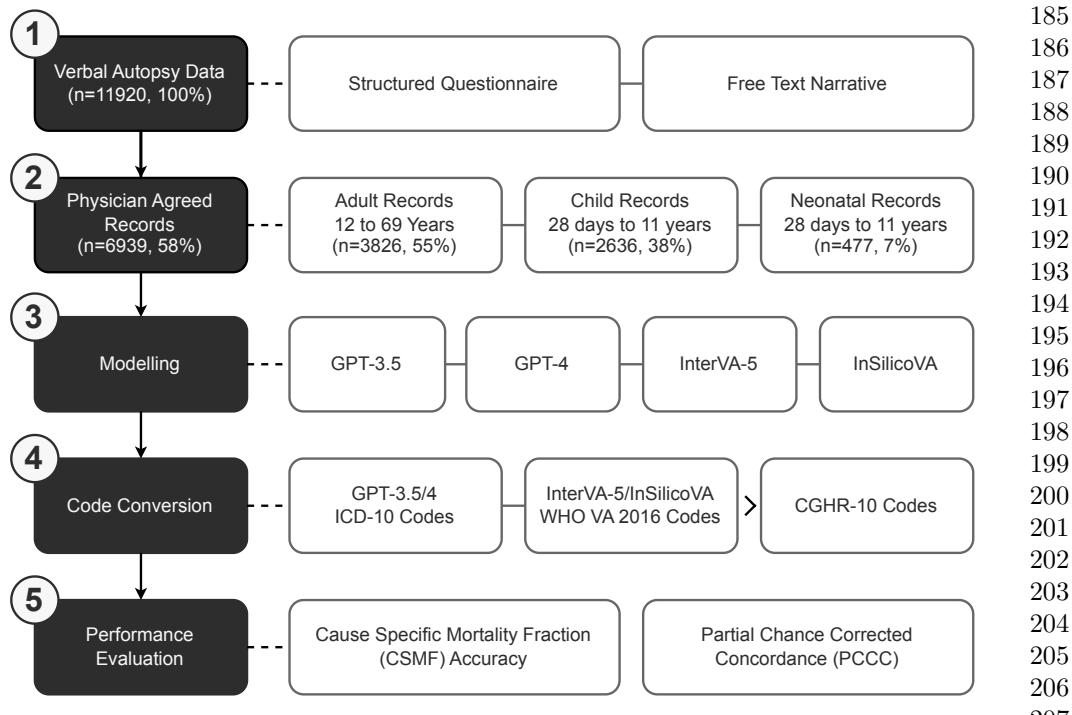


Fig. 1 Case study methods.

(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.

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]. Thus, we used 6942 physician agreed records of the 11,920 total records. For better comparison, we standardized all codes to CGHR-10 codes (see Additional File 1) that generalized ICD-10 codes into

231 19, 10, and 7 categories for the adult (12 to 69 years), child (28 days to 11 years), and
232 neonatal (under 28 days) age groups. After conversion, a final total of 6939 physician
233 agreed records (3826 adult, 2636 child, and 477 neonatal) were used for modelling and
234 performance evaluation. See Appendix A.1 for further details on data preprocessing
235 and Tables A1 and A2 for COD and age range distributions of the physician agreed
236 records.

240

241

242 2.2 Modelling

243

244 Four computer models were used to assign COD for each of the 6939 physician agreed
245 records: GPT-3.5, GPT-4, InterVA-5, and InSilicoVA. InterVA-5 and InSilicoVA are
246 widely used and studied standard statistical models [13, 21, 22, 24, 25, 41, 42] for COD
247 assignment in VAs under the openVA framework [43]. InterVA-5 applies Bayesian prob-
248 abilistic modelling [44] using a set of standardized symptoms from reports and related
249 conditional probabilities from medical experts to assign CODs based on the highest
250 probability [20, 45]. InSilicoVA improves upon InterVA (e.g. comparable probabilities
251 across individuals, measures of uncertainty, and inclusion of additional data sources)
252 with a hierarchical Bayesian framework and Markov Chain Monte Carlo (MCMC)
253 simulations [46–48] to incorporate multiple sources of uncertainty for assigning CODs
254 based on the highest probability [17]. GPT-3.5 [49] and GPT-4 [34] are LLMs that
255 utilize deep neural networks with transformer architectures [50] and reinforcement
256 learning from human feedback [51–54] to follow instructions from prompts and pro-
257 vide human-level responses, with known differences in GPT-4 possessing multimodal
258 capabilites (e.g. image/voice input/output), more recent training data, and improved
259 responses compared to ChatGPT-3 [33].

260

261 For GPT-3.5 and GPT-4, the following user prompt was used to instruct each
262 model to produce COD assignments as ICD-10 codes, where <age> and <sex> from
263 the questionnaire, and <narrative> from the narratives, were replaced with values
264 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>	277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322
For InterVA-5 and InSilicoVA, the standardized questionnaire data from EVA were converted into OpenVA format [43], before being used as input for each model to produce COD assignments as WHO VA 2016 codes [55]. All model outputs were converted to CGHR-10 codes to evaluate performances of models for COD assignment relative to physicians. See Appendix A.2 for additional details regarding input parameters, output data, and code conversions for each model.	
2.3 Performance Evaluation	293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322
The performance of the four models were evaluated with metrics at the population and individual level by comparing their CGHR-10 COD outputs for 6939 records. Cause Specific Mortality Fraction (CSMF) accuracy was used to evaluate models on the population level (see Appendix A.3.1), while Partial Chance Corrected Concordance (PCCC) was used to evaluate models on the individual level (see Appendix A.3.2) [56]. Both CSMF accuracy and PCCC metrics are between 0 and 1 with 0 indicating low performance and 1 indicating perfect performance at the population and individual level respectively. As model performance can vary across ages and specific causes [41, 42, 57], the CSMF accuracy and PCCC metrics were compared for each model overall, by age group (adult, child, neonatal), by CGHR-10 COD codes, and across age ranges. 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. See Appendix A.3 for more details on performance metrics and evaluation strategy for comparing each model.	

323 **3 Results**

324

325 This section details the performance results of GPT-3.5, GPT-4, InterVA-5, and InSil-
326 icoVA models for assigning CGHR-10 CODs after applying the methods in Section 2.
327 GPT-4 performed the best overall at 0.61 PCCC followed by GPT-3.5 at 0.56 PCCC.
328 GPT-4 also had the highest PCCC for most age ranges and CODs across the adult (12
329 to 69 years), child (28 days to 11 years), and neonatal (under 28 days) age groups with
330 GPT-3.5, InterVA-5, and InSilicoVA having higher PCCC values for a few age ranges
331 and CODs. Overall performance results are seen in Section 3.1, and performance by
332 adult, child, and neonatal records are seen in Sections 3.2, 3.3, and 3.4 respectively.
333

334 **3.1 Overall Performance**

335

336 Of all 6939 records, GPT-4 (0.61 PCCC) had the highest individual performance
337 followed by GPT-3.5 (0.56 PCCC), InSilicoVA (0.44 PCCC), and InterVA-5 (0.44
338 PCCC) (Figure 2). GPT-3.5 and GPT-4 had improvements ranging from 0.14-0.18
339 PCCC over InSilicoVA and InterVA-5, while GPT-4 slightly improved over GPT-3.5
340 by 0.05 PCCC. Population level performances were similar for all models (0.74-0.79
341 CSMF). Figure 3 shows the PCCC performance across three age groups (adult, child,
342 and neonate). GPT-4 had the best individual performance for adult and neonatal
343 records (0.64 and 0.58 PCCC), while GPT-3.5 had the best performance for child
344 records (0.54 PCCC) with GPT-4 performing slightly worse (0.51 PCCC). InSilicoVA
345 and InterVA-5 performed the worse for adult and child records (≤ 0.5 PCCC), while
346 GPT-3.5 performed the worse for neonatal records (0.42 PCCC). Across age ranges,
347 all models followed a similar pattern in individual performance (Figure 4). PCCC
348 trended upwards for 1 month to 14 years ($\sim 0.1-0.75$), and downwards for ages 15 to
349 69 years ($\sim 0.7-0.35$). The highest and lowest performances were observed for ages 10-
350 29 years (\sim) and 1-11 months respectively. Performances varied more across models
351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

for ages 0 days to 5 years, while less variation was seen between GPT-3.5 and GPT-4, 369
as well as InSilicoVA and InterVA-5, from 5 to 69 years. 370
371
372
373

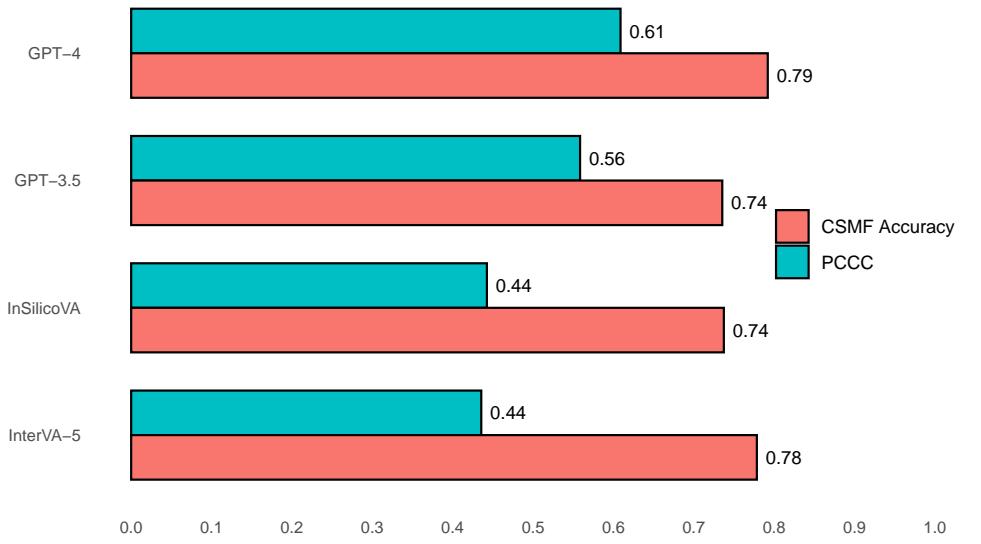


Fig. 2 Overall model performance.

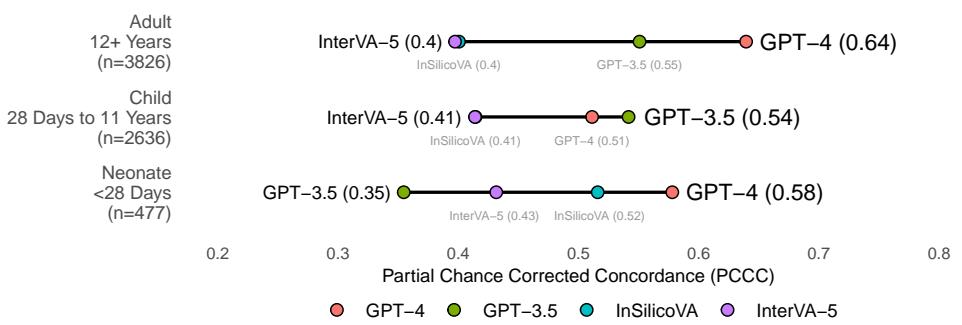


Fig. 3 Model performance by age group.

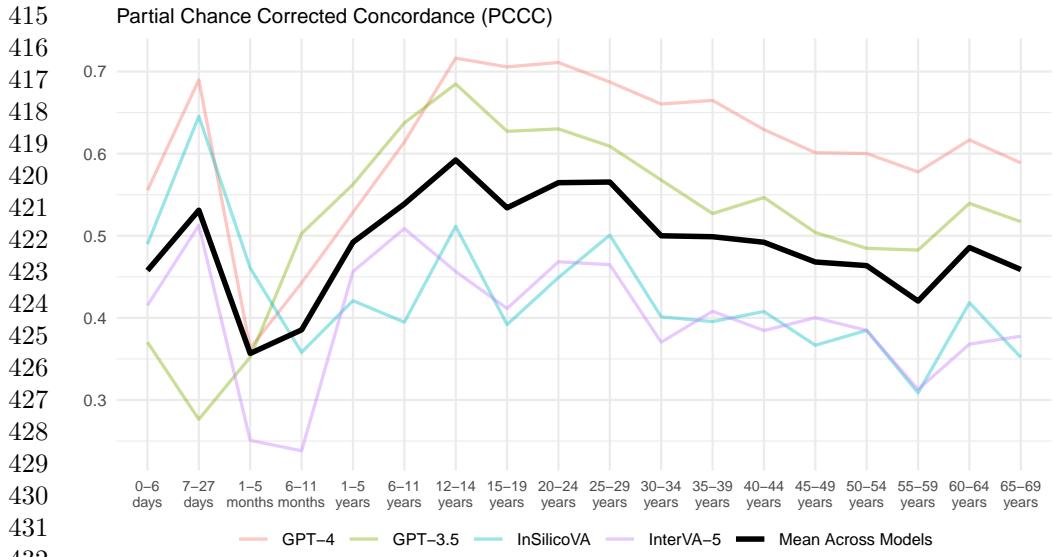


Fig. 4 Model performance by age range.

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

Figure 5 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

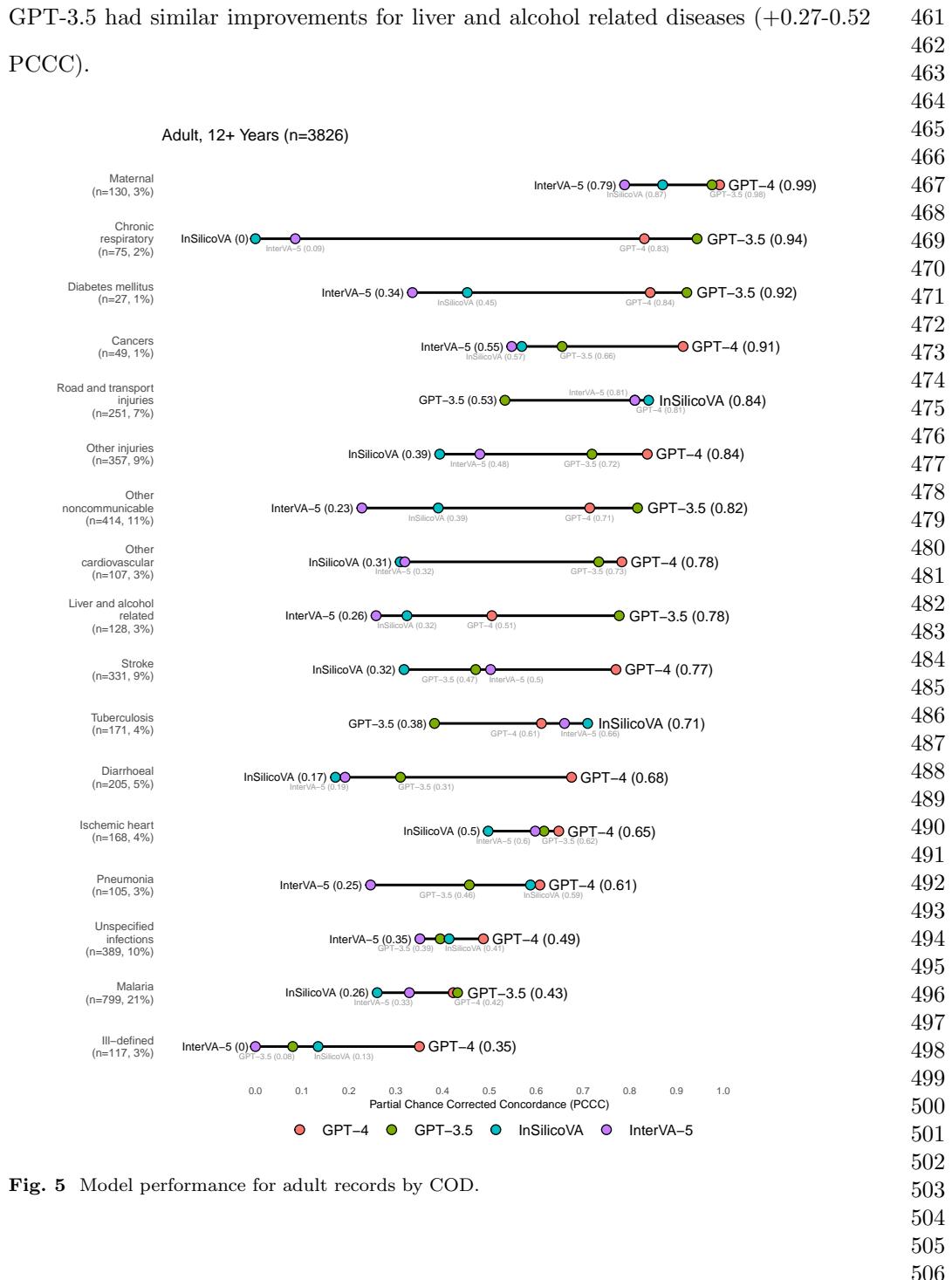


Fig. 5 Model performance for adult records by COD.

507 **3.3 Performance for 2636 Child Records (28 Days to 11 Years)**

508

509 Figure 6 presents individual performances for each of the models by 8 child CODs,
510 excluding congenital anomalies due to low sample size (n=1, <1%). GPT-4 had the
512 highest individual performance for 4 of 8 CODs (0.65-0.94 PCCC), GPT-3.5 for 3
513 CODs (0.44-0.88 PCCC), and InSilicoVA for 1 COD (0.78 PCCC). InterVA-5 had
515 the lowest performance for 4 of 8 CODs (0.09-0.79 PCCC), InSilicoVA for 3 CODs
517 (0-0.35 PCCC), and GPT-3.5 for 1 COD (0.58 PCCC). All models had >0.7 PCCC
518 for injuries (0.79-0.94 PCCC), and <0.6 PCCC for malaria (0.35-0.54 PCCC) and
519 other infections (0.29-0.44 PCCC). GPT-4 had improvements >0.3 PCCC compared
522 to other models for ill-defined CODs (+0.38-0.65 PCCC), and larger improvements
523 over other models for injuries (+0.11-0.15 compared to +0.01-0.04 PCCC).

525

526

527 **3.4 Performance for 477 Neonatal Records (Under 28 Days)**

528

529 Model performance across 5 neonatal CODs, excluding congenital anomalies (n=2,
530 <1%) and other (n=5, 1%) due to small sample sizes is shown in Figure 7. GPT-4
532 had the highest individual performance for 3 of 5 CODs (0.39-0.71 PCCC), GPT-3.5
533 for 1 COD (0.57 PCCC), and InSilicoVA for 1 COD (0.86 PCCC). GPT-3.5 had the
535 lowest performance for 3 of 5 CODs (0-0.13 PCCC) and InterVA-5 for 2 CODs (0.01
537 and 0.48 PCCC). All models had similar performance for stillbirth deaths (0.48-0.57
538 PCCC), while only GPT-4 had a PCCC >0 PCCC. InSilicoVA had improvements
539 over all other models for neonatal infection deaths (+0.18-0.73 PCCC).

541

542

543

544 **4 Discussion**

545

546

547 This section discusses and summarizes the results from Section 3. Advantages and dis-
548 advantages of using GPT-3.5, GPT-4, InterVA-5, and InSilicoVA models for assigning
549 CODs are discussed in Sections 4.1 and 4.2. Limitations of the study are mentioned
550 in Section 4.3, while opportunities and future work are detailed in Section 4.4.

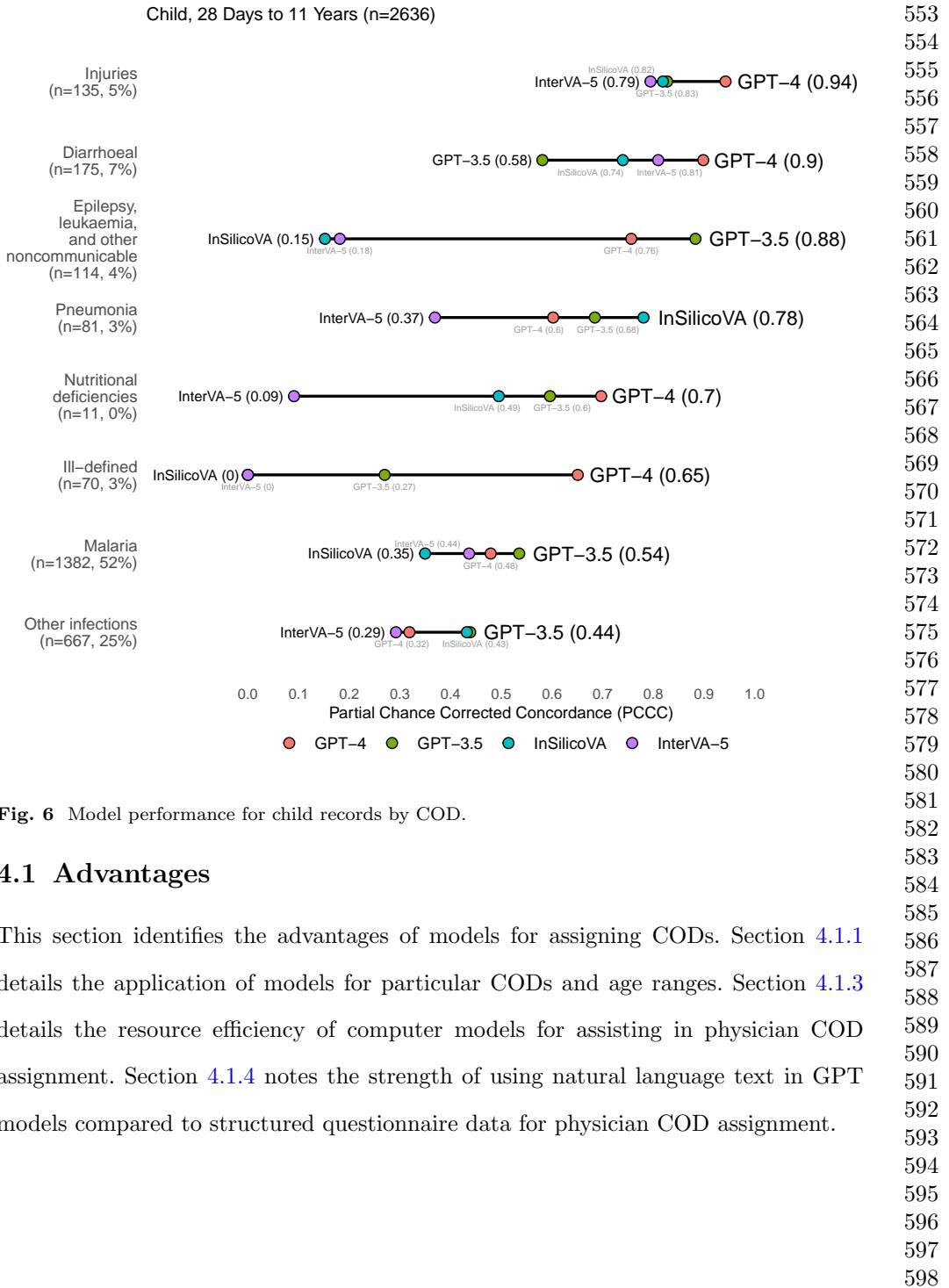


Fig. 6 Model performance for child records by COD.

4.1 Advantages

This section identifies the advantages of models for assigning CODs. Section 4.1.1 details the application of models for particular CODs and age ranges. Section 4.1.3 details the resource efficiency of computer models for assisting in physician COD assignment. Section 4.1.4 notes the strength of using natural language text in GPT models compared to structured questionnaire data for physician COD assignment.

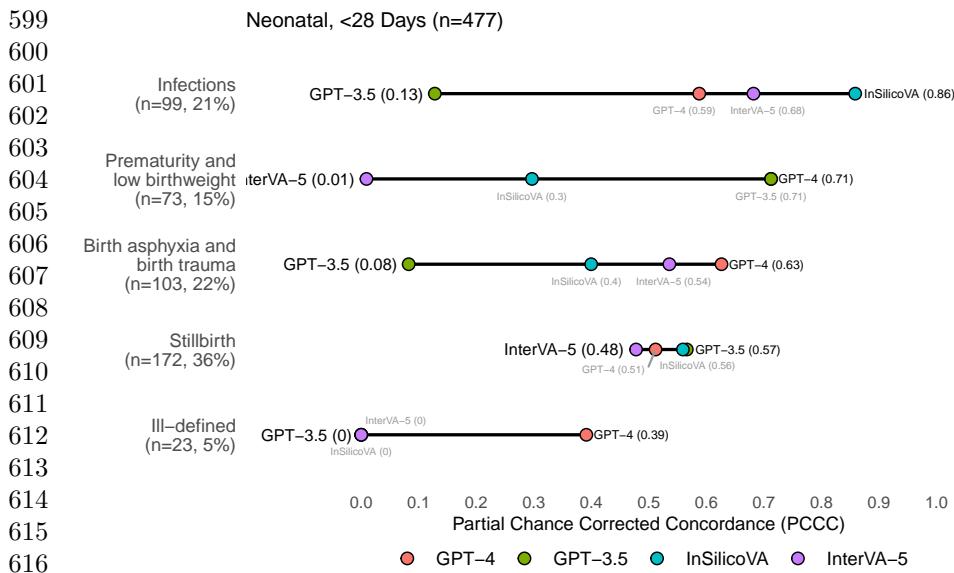


Fig. 7 Model performance for neonatal records by COD.

4.1.1 Cause-specific Models

At the population level, overall performances for all models were similar to physicians (0.74-0.79 CSMF), indicating potential for adequately estimating COD distributions for large populations. Although all models did not perform well for all records at the individual level (0.44-0.61 PCCC), several models performed well for certain CODs (0-0.99 PCCC). For most CODs, GPT-3.5/GPT-4 performed better than InSilicoVA/InterVA-5 (top PCCC for 15 of 17, 7 of 8, and 4 of 5 adult, child, and neonatal CODs respectively), while InSilicoVA performed better for particular CODs (road and transport injuries, tuberculosis, pneumonia, and neonatal infections with 0.84, 0.71, 0.78, and 0.86 PCCC respectively). For CODs with high performance (e.g. GPT-3.5/4 with 0.91-0.99 PCCC for maternal conditions, chronic respiratory disease, diabetes melitus, and cancers, InSilicoVA with 0.84 and 0.86 PCCC for road and transport injuries, and neonatal infections), the results suggest that GPT-3.5/4 and

InSilicoVA may assign CODs that are very similar to physicians. Thus, it may be beneficial to evaluate performance at the COD level, and apply a combination of models that perform well in comparison to physicians for each COD. For example, different models perform well for various leading CODs as seen in Table 1 [58, 59].

Table 1 Leading causes globally in 2021 and most relevant models.

Leading Causes of Death (~53% of 68M deaths) ¹	Deaths (% of 68M) ²	Best Model(s)	PCCC
Ischaemic heart disease	9M (13%)	GPT-4	0.65 (n=168)
Stroke	7M (10%)	GPT-4	0.77 (n=331)
Cancers	4.3M (4%)	GPT-4	0.91 (n=49)
Lower respiratory infections	2.4M (3%)	GPT-3.5/4	0.78 (n=180) ³
Diabetes mellitus	1.6M (2%)	GPT-3.5	0.92 (n=27)
Tuberculosis	1.4M (2%)	InSilicoVA	0.71 (n=171)
Hypertensive heart disease	1.4M (2%)	GPT-4	0.78 (n=107) ⁴
Cirrhosis of the liver	1.3M (2%)	GPT-3.5	0.78 (n=128) ⁵
Diarrhoeal diseases	1.2M (2%)	GPT-4	0.68 (n=205)
Road injury	1,183 (2%)	InSilicoVA	0.84 (n=357)
Preterm birth complications	0.9M (1%)	GPT-4	0.71 (n=73)
Falls	0.7M (1%)	GPT-4	0.89 (n=492) ⁶

¹COVID-19, kidney disease, alzheimer disease, other dementias, and self-harm were excluded as a relevant CGHR-10 code was not present. Trachea, bronchus, lung, colon, rectum, stomach, and breast cancers were generalized into cancers.

²Percentage of ~68 Million (M) deaths globally. Numbers are rounded.

³Mean of chronic and acute respiratory infections.

⁴Derived from other cardiovascular diseases.

⁵Derived from liver and alcohol related diseases.

⁶Mean of adult and child injuries.

4.1.2 Age-specific Performance Patterns

Across age ranges, all models followed a similar upward trend from 6 months to 14 years of age, and a downward trend from 15-69 years with GPT models having higher performance than InSilicoVA/InterVA-5 models, while more mixed trends were observed from 0 days to 5 months (recall Figure 4). For adult age ranges, performance generally decreased as age increased, which suggested that models had difficult assigning CODs for older than younger adults with some improvements after the age of 59. For child and

691 neonatal age ranges, the performance improved drastically as the age increased after 5
692 months, suggesting less difficulty in COD assignment when children and neonates are
693 more developed. As the models did not perform particularly well (≥ 0.8 PCCC) for
694 any specific five-year age range, it is not recommended to apply specific models that
695 target cases by age. However, the patterns of increases and decreases of performance in
696 relation to age provide valuable insight for comparison to expected physician diagno-
697 sis patterns in well-studied medical literature and knowledge. For example, it may be
698 expected that physicians are more uncertain in diagnosing diseases that are prevalent
699 in neonatal patients [60, 61], which are present in our findings from Figure 4.
700
701

702 4.1.3 Scalability and Availability

703

704 The models in this study can assist physicians in assigning CODs in a variety of ways
705 due to low costs and speed of COD assignment. Similar to differential diagnoses, GPT
706 and InSilicoVA models offer alternative COD assignments for physicians to consider
707 [39], which can potentially help lower the number of records with ill-defined causes or
708 reduce disagreement between physicians. At the time of this study, running GPT-3.5
709 cost $\sim \$1.6$ USD ($\$0.5$ per one million tokens), GPT-4 cost $\sim \$115$ USD ($\$30$ per one
710 million tokens), and InSilicoVA was cost free on 6939 records [62]. These costs were
711 lower than physicians (e.g. less than $\$3$ USD per house in India [15, 16]), while it is
712 possible to code over 10,000 records in under a day. When physicians are unavailable,
713 GPT and InSilicoVA models can be a cost-efficient alternative to code large amounts
714 of records for population estimates of CODs. However, it is recommended to apply
715 these models only for certain CODs where models perform well, such as in Table 1. In
716 addition, these models can also help divert physician resources to cases that are more
717 difficult to code or require more attention. For example, physicians can validate cases
718 where models performed well (e.g. maternal conditions at 0.79-0.99 PCCC), while
719 spending more time on cases where models performed poorly (e.g. acute respiratory
720 infections at 0.25-0.61 PCCC).
721
722

4.1.4 Natural Language Input and Output	737
	738
Training data was not required to assign CODs for all models, which allowed application without domain expertise or supplying training datasets. The main advantage to GPT-3.5/4 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 standard, 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 with other incompatible models. 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 [26, 28]. For example, GPT-3.5/4 had higher performance (+0.35-0.65 PCCC) than InterVA-5 and InSilicoVA for ambiguous ill-defined records across age groups. GPT models also performed better (+0.11-0.61 PCCC) on CODs with a rarer occurrence, such as nutritional deficiencies (n=11) and diabetes mellitus (n=27). Rarer CODs may be more difficult to capture by questionnaire due to lack of sample data, but it may possibly have richer contextual information from articles, web sources, or books that offer knowledge for GPT models to leverage.	739
	740
	741
	742
	743
	744
	745
	746
	747
	748
	749
	750
	751
	752
	753
	754
	755
	756
	757
	758
	759
	760
	761
	762
	763
	764
	765
	766
	767
	768
	769
	770
	771
	772
	773
	774
	775
	776
	777
	778
	779
	780
	781
	782
4.2 Disadvantages	773
This section discusses the disadvantages of GPT models for COD assignment. Section 4.2.1 identifies issues in reproducing GPT outputs for repeated runs on the same records and lack of up-to-date information, while Section 4.2.2 discusses the resource intensive infrastructure required by GPT and its relation to data privacy.	774
	775
	776
	777
	778
	779
	780
	781
	782

783 **4.2.1 Reproducibility and Timeliness**

784

785 Recall that the GPT models in this study had the temperature parameter set to 0 for
786 more reproducible and reliable results. A short experiment in Appendix B revealed
787 that GPT-3.5 assigns the same COD for the same record only more than 60% of the
788 time, based on repeated runs on a sample of 100 records. This suggests that GPT
789 models do not always reliably assign identical CODs for the same case on multiple runs,
790 which may pose issues in reproducibility and reliability. For example, GPT models
791 may achieve correct COD assignments solely due to random chance, but are difficult to
792 test with large numbers (e.g. 10,000) of reruns due to costs (e.g. costs increased 10 fold
793 per record when rerun 10 times). In comparison, InterVA-5 and InSilicoVA are open
794 source and free, allowing a large number of reruns without incurring additional fees.
795 In addition, InterVA-5 and InSilicoVA assign CODs and provide probabilities for each
796 alternative COD, which offers more reproducible and reliable COD assignments despite
797 lower performance overall. Lastly, a major disadvantage in all models was that they
798 were trained on historical data up to particular points in time, which may not utilize
799 the most up-to-date data available (e.g. latest online articles, social media, or books
800 for GPT models). Emergent diseases (e.g. COVID-19) and changes in distributions
801 (e.g. outbreaks) may not be caught by these models depending on how often they are
802 updated.

803

804 **4.2.2 Infrastructure and Data Privacy**

805

806 GPT-3.5 and GPT-4 models required large computing infrastructure to train and run,
807 which was not possible to run on local computers, or setup due to costs and ownership
808 of the models. This poses issues with data privacy as sensitive data (e.g. identifying
809 information) need to be sent to company servers, which can be collected by companies
810 (e.g. OpenAI) and misused [63]. For example, in our study, GPT models use prompts,
811 which contain the narrative data, to assign CODs, and the data in these prompts
812 813

may be unknowingly collected and misused by companies (e.g. companies) or their users (e.g. malicious prompts) to identify participants or leak sensitive sensitive data [64, 65]. 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.	829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874
4.3 Limitations	838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874
This section identifies limitations in this research in the context of GPT models. Section 4.3.1 identifies the omission of ICD-10 performance evaluations. Section 4.3.2 mentions the need for parameter tuning and evaluation of consistency and multiple COD assignments.	838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874
4.3.1 ICD-10 Evaluation and Low Sample Sizes	848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874
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 are often low and inadequate for training statistical models. In relation, some broader CGHR-10 CODs were even removed for performance evaluation as <10 cases were captured (e.g. congenital anomalies, suicide). Although GPT models were able to assign ICD-10 codes, lower performance may be expected as even physicians do not agree completely on ICD-10 codes, noted that broader categories (CMEA-10 codes in Additional file 2) were used to assign equivalency or agreement.	850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874
4.3.2 Model Tuning, Consistency, and Multiple Outputs	870 871 872 873 874
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	870 871 872 873 874

875 model settings may be adjusted to possibly improve performance for GPT models
876 [66]. This was not examined as sensitivity analyses on model parameters are costly
877 across multiple reruns, noted in Section 4.2.1, which is required when testing various
878 parameter settings. In addition, GPT models may possibly produce inconsistent results
881 even with the temperature set to 0. Thus, it is important to also test the reliability
882 and consistency of GPT outputs to avoid coincidental results due to randomness [67–
884 69]. InterVA-5 and InSilicoVA were able to provide multiple COD assignments with
886 probabilities for each COD. GPT models can be prompted to produce more than one
887 COD assignment, but was not explored in this study as we only evaluated the most
889 probable COD. This may be useful to evaluate the performance of multiple alternative
891 COD assignments, which may provide additional diagnoses that have a higher chance
893 of being similar to physician assignment, and better reflect causes leading to death
894 [19].

896

897

898 **4.4 Opportunities**

899

900 This section discusses research opportunities to improve GPT models for assigning
901 CODs. Section 4.4.1 discusses the potential to improve GPT models with prompt
903 engineering and exploration of misclassified records, while Section 4.4.2 describes the
905 application of GPT models for improving household surveys for better data quality.
906 Section 4.4.3 identifies an opportunity to integrate GPT, InterVA-5, and InSilicoVA
908 models into VA systems for improving physician COD assignment.

909

910

911 **4.4.1 Prompt Engineering and Custom Models**

912

913 Prompt engineering, the design of prompts to guide GPT models for better results [70],
914 presents an important research opportunity that may improve performance of GPT
915 models for COD assignment. An example exploration was conducted in Appendix C on
917 misclassified GPT-4 records for neonatal infections, which found potential issues with
919 920

the categorization of CGHR-10 codes, order of information in narratives, and guidelines of COD assignments. An analysis of misclassified records with domain experts (e.g. physicians, specialists) may yield insights on adjusting prompts to assign more correct CODs, or apply more relevant broad COD categories for evaluation. In addition, subsequent prompts, data, and examples can be used to include correctional instructions and refine results, while additional information from the questionnaire and physician VA manuals can provide contextual information (e.g. retrieval augmented generation [71]) for further performance improvements [72]. 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 (e.g. custom models for particular CODs) [73].	921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966
4.4.2 Guided and Monitored Household Surveys	944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966
Recall that VAs involve surveyors that visit households to gather information about the deceased from their family, next-of-kin, friend, or community. Although standard questionnaires are used during this visit, there is significant information, containing latent patterns, from the narrative that is not always captured by the questionnaire [26, 28]. These narratives often require a human connection between the surveyor and household members, where surveyor characteristics vary in social ability, cultural understanding, emotional capacity, and medical knowledge that affect the quality and bias of narratives [19, 74]. GPT models may help guide surveyors during VA interviews to probe households for better narrative information by generating and suggesting better questions, or providing questions that may have been missed by the surveyors. In addition, as models can assign CODs on-demand, there is potential for models to provide immediate COD estimates during the data collection process to monitor	944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966

967 data quality on-demand (e.g. comparing estimated to expected COD distributions for
968 known areas as quality checks).
969

970

971 **4.4.3 Computer Assisted Verbal Autopsy** 972

973 Our study lays the foundation for the integration of GPT, InterVA-5, and InSilicoVA
974 models into VA systems to assist physicians in COD assignment. In dual-coded VA sys-
975 tems (described in Section 2.1), two physicians are randomly assigned to each record
976 and require second inspections of each other’s assignment (reconciliation) and evalua-
977 tion by a third more senior physician if their assignments do not agree. As mentioned
978 in Section 4.1.3, suggestion of alternative assignments from GPT and InSilicoVA mod-
979 els potentially reduces the disagreement between physicians, and ill-defined records,
980 while allowing physicians to focus on more difficult records. Thus, model suggestions
981 can be integrated into VA systems by presenting COD suggestions to physicians after
982 their initial COD assignment, which allows them to consider alternative assignments
983 and possibly revise their assignments based on the suggestions. At step 2 in Figure 8,
984 GPT, InterVA-5, and InSilicoVA models can suggest COD assignments to consider,
985 providing the option in step 2b to revise or proceed with their initial assignment.
986 Our future work will be a first step in computer assisted verbal autopsy, assessing the
987 effects of these model suggestions on improve VA data quality (e.g. increase in agreed
988 records, reduction of ill-defined deaths). In preparation, we have integrated GPT-
989 4, InterVA-5, and InSilicoVA model suggestions into our on-going HEAL-SL study
1000 1001 after survey round 2 [35] with goals of increasing physician agreement and reducing
1002 1003 ill-defined COD assignments.
1004

1005

1006

1007 **5 Conclusion** 1008

1009 This study evaluates the performance of GPT-3.5, GPT-4, InterVA-5, and InSilicoVA
1010 1011 models compared to physicians for assigning CODs for 6939 VA records in Sierra
1012

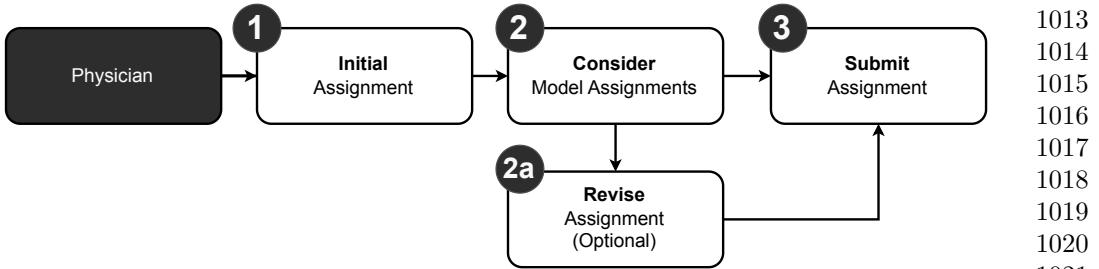


Fig. 8 Model suggestions integrated in the physician assignment process.

Leone (2019-2022). At the population level, all models were similar (0.74-0.79 CSMF accuracy). At the individual level, GPT-4 had the best performance (0.61 PCCC), followed by GPT-3.5 (0.58 PCCC), and InSilicoVA/InterVA-5 (0.44 PCCC). Across CODs, GPT-4 had performed best for 10 of 17 adult, 4 of 8 child, and 3 of 5 neonatal CODs, with GPT-3.5 for 5 adult, 3 child, and one neonatal CODs, and InSilicoVA for 2 adult, one child, and one neonatal CODs. Model performance increased (\sim 0.1-0.75 PCCC) as children and neonates developed (0 days to 14 years), and decreased (\sim 0.7-0.35) as adults aged (15 to 69 years). Thus, GPT and InSilicoVA models were comparable to physicians for several CODs, but not across age ranges. As performance varied across CODs and age ranges, it is advantageous to combine several models to target CODs that each model performs well for, and to compare age-related performance patterns in relation to physicians. In addition, all models were able to scale to a large number of records and were available on-demand in comparison to physicians, enabling COD estimation and alternative diagnoses in low resource or physician scarce scenarios. As GPT models operate on natural language, they are able to adapt to more loosely defined data structures (e.g. assign in different COD coding standards, provide reasoning, and use contextual information when samples are low), making them behave more similarly to physician assignment. However, GPT models do not provide reliable CODs on repeated assignment, and were limited to past training data, with large computing infrastructure requirements, leading to reproducibility issues in COD

1059 assignments, difficulty adapting to new or changing CODs, and data privacy issues.
1060
1061 Limitations of this study included difficulty comparing ICD-10 codes directly due to
1062 incompatible COD outputs from each model and low sample sizes, difficulty in con-
1063 ducting sensitivity analyses for GPT models due to costs, and omitting evaluation of
1064 multiple COD assignments due to study scope. We identified research opportunities
1065 in refining GPT models using prompt engineering and custom models for improving
1066 performance, guided household surveys to improve narrative quality, and future work
1067 in computer assisted VA, where GPT and other models will be used to assist physician
1068 COD assignment by offering multiple alternative assignments, with goals of increasing
1069 agreement on COD assignment and reducing ill-defined deaths. GPT-4, InterVA-5,
1070 and InSilicoVA has been integrated into future survey rounds of the HEAl-SL study
1071 from 2022 onwards, offering alternative COD assignments to assist physicians with
1072 second opinions. Future work in evaluating the effectiveness of computer assisted VA
1073 to reduce disagreements among physicians and ill-defined deaths will help support the
1074 advancement of more accurate and efficient VA systems across the world.
1075
1076

1077 **Supplementary information.** Additional files were used to supplement this paper:
1078

- 1079 • Additional file 1: Centre for Global Health Research 10 (CGHR-10) codes. Codes
1080 grouping ICD-10 code ranges into generalized categories. (.csv)
- 1081 • Additional file 2: Central Medical Evaluation Agreement 10 (CMEA-10) codes. ICD-
1082 10 code ranges considered in physician agreement. (.csv)

1083
1084 **Acknowledgments.** TBD.

1085
1086

1087 **Declarations**

1088
1089

1090 **Funding**

1091
1092 TBD.

1093
1094

Competing interests	1105
Not applicable.	1106
Ethics approval	1107
Not applicable.	1108
Consent for publication	1109
Not applicable.	1110
Availability of data and materials	1111
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) and narra- tive data by age group and survey rounds 1 and 2 available at https://openmortality.org/dataset/heal-sl . Cause of death code mappings to convert between ICD-10, WVA-2016, and CGHR-10 codes available at https://openmortality.org/dataset/icd . Model evaluation result files at https://github.com/cghr-toronto/healsl-gpt-paper/tree/main/data .	1112
Code availability	1113
All code for this paper is available at https://github.com/cghr-toronto/healsl-gpt-paper .	1114
Authors' contributions	1115
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.	1116

1151 AB and RCM provided medical domain guidance and feedback. All authors reviewed
1152 the results and contributed to the report. All authors read and approved the final
1153 manuscript.
1154

1155

1156

1157 **Appendix A Details on Methods**

1158

1159
1160 This section provides additional details on the methods described in Section 2. An
1161 overview of the methods used in this study is seen in Figure A1 as a five-step process.
1162 Section A.1 provides details on the preprocessed data used for modelling. Section A.2
1163 describes the data and parameter inputs and outputs for each model, while Section
1164
1165 A.3 details the evaluation of model outputs at the individual and population level
1166 across different CODs, age groups, and age ranges.
1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

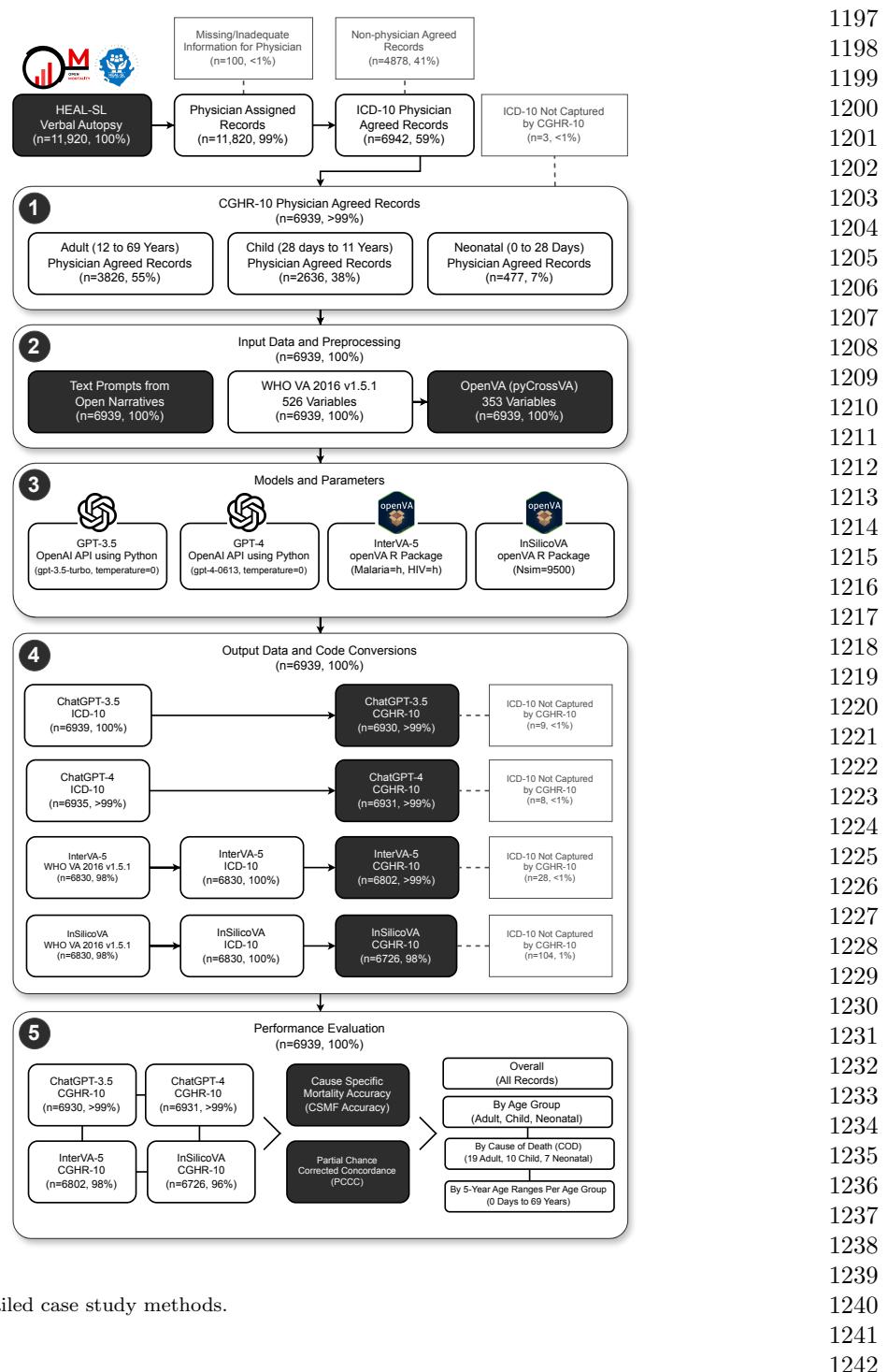


Fig. A1 Detailed case study methods.

1243 **A.1 CGHR-10 Physician Agreed Records**

1244

1245 Initially, 11,920 records were collected from dual-coded EVA in the HEAL-SL study.
1246
1247 Physicians were able to assign CODs for 11,820 of the 11,920 records, where 100 of
1248 these records could not be assigned a COD due to missing or inadequate information
1249
1250 (e.g. low quality narrative, data loss). The 11,820 physician coded records were further
1251
1252 filtered for records where both physicians agreed on the assigned codes (records that
1253 were not reconciled or adjudicated) resulting in 6942 physician agreed records (based
1254
1255 on comparisons using CMEA-10 codes, see Additional File 2). The 6942 records were
1256
1257 converted into CGHR-10 codes (see Additional File 1) that generalized ICD-10 codes
1258 into 19, 10, and 7 categories for the adult (12 to 69 years), child (28 days to 11
1259 years), and neonatal (under 28 days) age groups. After conversion, a final total of
1260
1261 6939 physician agreed records (3826 adult, 2636 child, and 477 neonatal) were used
1262
1263 for modelling and performance evaluation, where three records were removed as their
1264
1265 ICD-10 codes did not have a matching CGHR-10 code.

1266

1267 The 6939 physician agreed records were collected using VA from the HEAL-SL
1268 study between 2019-2022, where records were collected using nation wide samples
1269 across Sierra Leone provinces seen in Figure A2. More populous areas (e.g. southern
1270
1271 and north east provinces with ~197,000 and ~135,000 population respectively) had
1272 more sampling areas versus less populous areas (e.g. north west and eastern provinces
1273 with ~50,000 and ~69,000 people respectively). The distribution of the case study
1274
1275 data are shown by CGHR-10 causes of death in Table A1. All age groups had relatively
1276
1277 evenly distributed female and male records (44-55% of 6939 records each). Across
1278
1279 CODs, there were noticeably more female records for cancers (65%), and maternal con-
1280 ditions (100%), while more male records for chronic respiratory diseases (61%), other
1281
1282 noncommunicable diseases (61%), other injuries (77%), road and transport injuries
1283 (71%), and tuberculosis (68%). Most records were coded by physicians as malaria for
1284 adults (20%) and children (52%), and stillbirth (36%) and neonatal infections (21%)
1285
1286

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

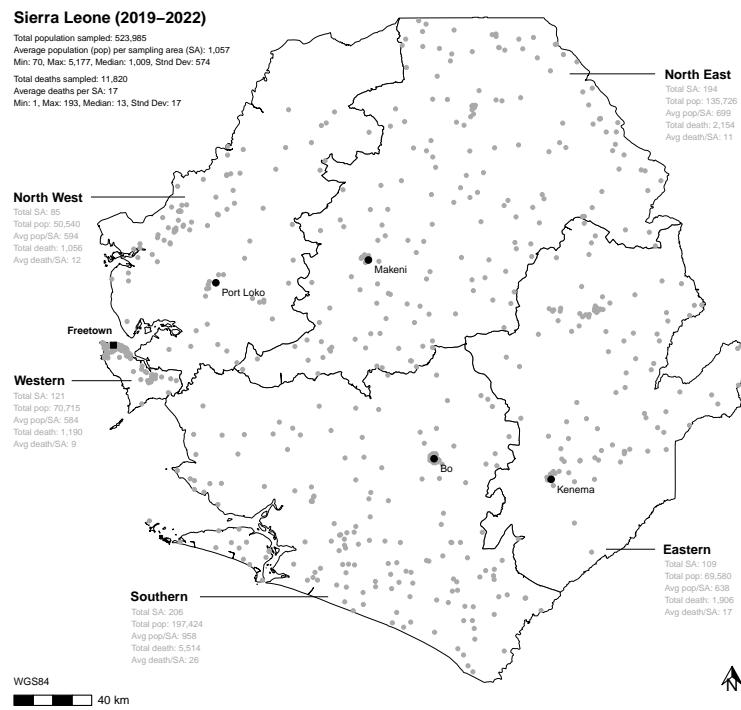


Fig. A2 Case study data sampling areas.

A.2 Modelling Details

Each model (GPT-3.5, GPT-4, InSilicoVA, and InterVA-5) required pre-processing of the 6939 records into input data, and standardization of output COD codes from models for performance evaluation as not all models produced comparable codes across outputs. Although each model can assign multiple CODs per record, only the first

1335

1336 **Table A1** Case study data by cause of death.

1337 1338	Age Group	CGHR-10 Cause of Death (COD)	Female	Male	Total
1339		Acute Respiratory Infections	48 (45.7%)	57 (54.3%)	105 (2.7%)
1340		Cancers	32 (65.3%)	17 (34.7%)	49 (1.3%)
1341		Chronic Respiratory Diseases	29 (38.7%)	46 (61.3%)	75 (2%)
1342		Diabetes Mellitus	14 (51.9%)	13 (48.1%)	27 (0.7%)
1343		Diarrhoeal Diseases	102 (49.8%)	103 (50.2%)	205 (5.4%)
1344	Adult, 18 CODs (n=3826, 55.1%)	Ill-Defined	56 (47.9%)	61 (52.1%)	117 (3.1%)
1345	Adult Female (n=1681, 43.9%)	Ischemic Heart Disease	89 (53%)	79 (47%)	168 (4.4%)
1346	Adult Male (n=2145, 56.1%)	Liver And Alcohol Related Diseases	58 (45.3%)	70 (54.7%)	128 (3.3%)
1347		Malaria	372 (46.6%)	427 (53.4%)	799 (20.9%)
1348		Maternal Conditions	130 (100%)	N/A	130 (3.4%)
1349		Other Cardiovascular Diseases	59 (55.1%)	48 (44.9%)	107 (2.8%)
1350		Other Noncommunicable Diseases	160 (38.6%)	254 (61.4%)	414 (10.8%)
1351		Other Injuries	83 (23.2%)	274 (76.8%)	357 (9.3%)
1352		Road And Transport Injuries	73 (29.1%)	178 (70.9%)	251 (6.6%)
1353		Stroke	147 (44.4%)	184 (55.6%)	331 (8.7%)
1354		Suicide	N/A	3 (100%)	3 (0.1%)
1355	Child, 9 CODs (n=2636, 38%)	Tuberculosis	54 (31.6%)	117 (68.4%)	171 (4.5%)
1356	Child Female (n=1290, 48.9%)	Unspecified Infections	175 (45%)	214 (55%)	389 (10.2%)
1357		Congenital Anomalies	1 (100%)	N/A	1 (0%)
1358		Diarrhoeal Diseases	79 (45.1%)	96 (54.9%)	175 (6.6%)
1359	Child Male (n=1346, 51.1%)	Epilepsy, Leukaemia, And Other Noncommunicable Diseases	61 (53.5%)	53 (46.5%)	114 (4.3%)
1360		Ill-Defined	34 (48.6%)	36 (51.4%)	70 (2.7%)
1361		Injuries	51 (37.8%)	84 (62.2%)	135 (5.1%)
1362	Neonate, 7 CODs (n=477, 6.9%)	Malaria	680 (49.2%)	702 (50.8%)	1382 (52.4%)
1363		Nutritional Deficiencies	7 (63.6%)	4 (36.4%)	11 (0.4%)
1364	Neonate Female (n=227, 47.6%)	Other Infections	338 (50.7%)	329 (49.3%)	667 (25.3%)
1365	Neonate Male (n=250, 52.4%)	Pneumonia	39 (48.1%)	42 (51.9%)	81 (3.1%)
1366		Birth Asphyxia And Birth Trauma	38 (36.9%)	65 (63.1%)	103 (21.6%)
1367		Congenital Anomalies	2 (100%)	N/A	2 (0.4%)
1368		Ill-Defined	11 (47.8%)	12 (52.2%)	23 (4.8%)
1369		Neonatal Infections	49 (49.5%)	50 (50.5%)	99 (20.8%)
1370		Other	2 (40%)	3 (60%)	5 (1%)
1371		Prematurity And Low Birthweight	39 (53.4%)	34 (46.6%)	73 (15.3%)
1372		Stillbirth	86 (50%)	86 (50%)	172 (36.1%)

1368

1369 generated COD response from GPT-3.5 and GPT-4, and the most probable COD
 1370 from InterVA-5 and InSilicoVA were used for evaluation. Section [A.2.1](#) describes the
 1371 input data and parameters for each model, while Section [A.2.3](#) details the outputs
 1372
 1373 from running each model.
 1374

1375

1376

1377

1378

1379

1380

Table A2 Case study data by age range.

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

A.2.1 Input Data and Preprocessing

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

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

User prompts contained textual instructions to perform coding of VA records based on the age, sex, and narrative of the deceased. The following template was used to

1427 generate user prompts for each record, where <age> and <sex> from the questionnaire,
1428
1429 and <narrative> from the narratives, were replaced with values from the data:
1430
1431 Determine the underlying cause of death and provide the most
1432 probable ICD–10 code for a verbal autopsy narrative of a <age>
1433
1434 years old <sex> death in Sierra Leone: <narrative>
1435
1436 For InterVA-5 and InSilicoVA, the standardized questionnaire data from the HEAL-SL
1437 EVA were first converted into 2016 World Health Organization (WHO) VA question-
1438 naire revision 1.5.1 Open Data Kit (ODK) format [75, 76] consisting of 526 variables
1439
1440
1441 [77], followed by further conversion into OpenVA format [43] consisting of 353 vari-
1442 ables [78] using the pyCrossVA version 0.97 Python package [79]. The 6939 records
1443
1444 were all converted into OpenVA formatted records for InterVA-5 and InSilicoVA.
1445
1446

1447 **A.2.2 Models and Parameters**

1448

1449 The GPT-3.5 and GPT-4 Application Programming Interface (API) was accessed
1450
1451 using Python version 3.11.4 and used to assign CODs for each record. GPT-3.5 used
1452 the gpt-3.5-turbo model, while GPT-4 used the gpt-4-0613 model. The parameter
1453 temperature for GPT-3.5 and GPT-4, representing the sampling temperature ranging
1454
1455 from 0 to 2 (default of 1), was set to 0 to produce more deterministic outputs [66].
1456
1457 Higher values closer to 2 may produce less deterministic outputs, while lower values
1458
1459 closer to 0 produce more deterministic outputs.

1460

1461 The openVA R package was used to run InterVA-5 and InSilicoVA models to assign
1462 CODs for each record in R version 4.3.1. The openVA package version 1.1.1 used
1463 dependent packages InterVA5 version 1.1.3 and InSilicoVA version 1.4.0. The Nsim
1464
1465 (number of iterations to run) parameter [80] for InSilicoVA was set to 9500, while
1466
1467 the HIV (level of prevalence of human immunodeficiency virus) and Malaria (level
1468
1469 of prevalence of Malaria) parameters [81] for InterVA-5 were both set to 'h' (high)
1470
1471 reflecting HIV and Malaria disease assumptions in Sierra Leone [82, 83]. Note that the
1472

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. 1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500

A.2.3 Output Data and Code Conversion

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

A.3 Performance Evaluation Details

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

1519 **A.3.1 Cause Specific Mortality Fraction (CSMF) Accuracy**

1520

1521 CSMF accuracy measures the performance of models at the population level, compar-
1522 ing distributions of CODs between the physicians and the models [56]. To calculate
1523 CSMF accuracy, we first calculate $CSMF_j$ as is the fraction of physician or model
1524 records for cause j , given by dividing the number of records for cause j with the total
1525 number of records as seen in Equation A1. Then, the $CSMFMaximumError$, repre-
1526 senting the worst possible model, is calculated using Equation A2. Finally, the CSMF
1527 accuracy is given by Equation A3, where k is the number of causes, j is a cause,
1528 $CSMF_j^{true}$ is the true physician CSMF for cause j , and $CSMF_j^{pred}$ is the prediction
1529 model CSMF for cause j . CSMF accuracy ranges from 0 to 1, where 1 means that the
1530 model completely matched the physician COD distribution and 0 means that it did
1531 not match the distribution at all.

1532

1533

1534

1535

$$CSMF_j = Records_j / Records \quad (A1)$$

1536

1537

1538

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

1539

1540

1541

1542

1543

1544

1545

1546

1547

1548

1549

1550

1551

1552

A.3.2 Partial Chance Corrected Concordance (PCCC)

1553

1554

1555

1556

1557

1558

1559

1560

1561

1562

1563

1564

PCCC measures the performance of models at the individual level, comparing COD assignments between the physicians and models on a record by record basis, correcting for COD assignments made purely by chance [56]. PCCC is given by Equation A5,

where k is the number of top COD assignments from the model to consider, N is

number of causes, and C is fraction of records where the physician COD assignment is

one of the top COD assignments from the model. For this study, 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 A4. Higher PCCC values closer to 1 indicate that model COD assignments are similar to physician COD assignments, while values closer to 0 indicate that model COD assignments are not similar to physicians.

$$C = \frac{TP}{Records} \quad (\text{A4})$$

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

Appendix B Experiment on Repeated Runs of GPT-3.5

A short experiment was conducted to test the consistency of GPT-3.5 outputs repeated on the same record. 100 records, sampled randomly with approximately equal proportions across age groups, CODs, and survey rounds 1 and 2, were used to test repeated runs of GPT-3.5. Each record from the 100 records was rerun 10 times through GPT-3.5, resulting in ten COD outputs per record. The ICD-10 codes were then converted to CGHR-10 codes and tested for consistency, where completely inconsistent results had different ICD-10 or CGHR-10 codes for each of the 10 reruns (1 times+), and completely consistent results had the same ICD-10 or CGHR-10 code for all 10 reruns (10 times), on the same record.

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

1611 to 0, but does so for more than half the records. For most records (more than 90%),
1612 GPT-3.5 will produce the same outputs for the same record 7 times or more out of 10.
1613

1614

1615 **Table B3** Records with same GPT-3.5 outputs based on 10 repeated
1616 reruns of 100 records

	Times with Same GPT-3.5 Outputs	ICD-10 Records	CGHR-10 Records
1619	1 times+ (inconsistent)	100	100
1620	2 times+	100	100
1621	3 times+	100	100
1622	4 times+	100	100
1623	5 times+	100	100
1624	6 times+	94	96
1625	7 times+	92	94
1626	8 times+	86	91
1626	9 times+	79	86
1626	10 times (consistent)	66	79

1627

1628

1629

1630

1631 **Appendix C Exploration of Neonatal Infections**

1632

1633 An exploration of neonatal infections (n=99, 21% of 477 records) was done to under-
1634 stand the low performance of GPT models (0.23 PCCC) for neonatal infections, and
1635 high performance of InSilicoVA (0.87 PCCC). In Table C4, about half the records
1636 were assigned correctly, and a majority (n=33, 33%) of the other records were mis-
1637 classified as other, while prematurity and low birthweight, birth asphyxia & birth
1638 trauma, and ill-defined make up the rest. On closer inspection of the 49 records with
1639 misclassified assignments, the ICD-10 code R50 was assigned in 20 records. R50 falls
1640 under unspecified infections in the adult CGHR-10 category, but in the other cate-
1641 gory for neonates. B50 was assigned in 4 records, falling under malaria, but a similar
1642 B54 falls under neonatal infections. P81 was assigned in 3 records, referring to fever
1643 of unknown origin, which falls under other, and P07 was assigned in 7 records, falling
1644 under prematurity and low birthweight.

1645

1646 In most misclassified records, there is mention of infections, but the misclassifica-
1647 tions occur due to the finer details of the ICD-10 code classifications, the categorization
1648

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

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

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

1703 **References**

- 1704
- 1705 [1] World Health Organization.: Non Communicable Diseases: Key Facts.
1706
1707 https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.
1708
- 1709 [2] Benziger CP, Roth GA, Moran AE. The Global Burden of Disease Study and
1710 the Preventable Burden of NCD. Global Heart. 2016 Dec;11(4):393–397. <https://doi.org/10.1016/j.gheart.2016.10.024>.
- 1711
- 1712
- 1713
- 1714
- 1715 [3] Lawn JE, Kerber K, Enweronu-Laryea C, Cousens S. 3.6 Million Neonatal
1716 Deaths—What Is Progressing and What Is Not? Seminars in Perinatology. 2010
1717 Dec;34(6):371–386. <https://doi.org/10.1053/j.semperi.2010.09.011>.
- 1718
- 1719
- 1720
- 1721 [4] Lassi ZS, Bhutta ZA. Community-based Intervention Packages for Reducing
1722 Maternal and Neonatal Morbidity and Mortality and Improving Neonatal Out-
1723 comes. Cochrane Database of Systematic Reviews. 2015;(3). <https://doi.org/10.1002/14651858.CD007754.pub3>.
- 1724
- 1725
- 1726
- 1727
- 1728 [5] Liu NH, Daumit GL, Dua T, Aquila R, Charlson F, Cuijpers P, et al. Excess
1729 Mortality in Persons with Severe Mental Disorders: A Multilevel Intervention
1730 Framework and Priorities for Clinical Practice, Policy and Research Agendas.
1731
1732 World Psychiatry. 2017;16(1):30–40. <https://doi.org/10.1002/wps.20384>.
- 1733
- 1734
- 1735
- 1736 [6] Ewig S, Torres A. Community-Acquired Pneumonia as an Emergency: Time for
1737 an Aggressive Intervention to Lower Mortality. European Respiratory Journal.
1738 2011 Aug;38(2):253–260. <https://doi.org/10.1183/09031936.00199810>.
- 1739
- 1740
- 1741
- 1742 [7] World Health Organization. SCORE for Health Data Technical Package: Global
1743 Report on Health Data Systems and Capacity, 2020; 2021.
- 1744
- 1745
- 1746
- 1747
- 1748

- [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>. 1749
 1750
 1751
 1752
 1753
 1754
 1755
 1756
 1757
 1758
 1759
 1760
 1761
 1762
 1763
 1764
 1765
 1766
 1767
 1768
 1769
 1770
 1771
 1772
 1773
 1774
 1775
 1776
 1777
 1778
 1779
 1780
 1781
 1782
 1783
 1784
 1785
 1786
 1787
 1788
 1789
 1790
 1791
 1792
 1793
 1794
- [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>. 1756
 1757
 1758
 1759
 1760
 1761
 1762
 1763
 1764
 1765
 1766
 1767
 1768
 1769
 1770
 1771
 1772
 1773
 1774
 1775
 1776
 1777
 1778
 1779
 1780
 1781
 1782
 1783
 1784
 1785
 1786
 1787
 1788
 1789
 1790
 1791
 1792
 1793
 1794
- [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>. 1762
 1763
 1764
 1765
 1766
 1767
 1768
 1769
 1770
 1771
 1772
 1773
 1774
 1775
 1776
 1777
 1778
 1779
 1780
 1781
 1782
 1783
 1784
 1785
 1786
 1787
 1788
 1789
 1790
 1791
 1792
 1793
 1794
- [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>. 1769
 1770
 1771
 1772
 1773
 1774
 1775
 1776
 1777
 1778
 1779
 1780
 1781
 1782
 1783
 1784
 1785
 1786
 1787
 1788
 1789
 1790
 1791
 1792
 1793
 1794
- [12] World Health Organization. Verbal Autopsy Standards: 2022 WHO Verbal Autopsy Instrument; 2023. 1775
 1776
 1777
 1778
 1779
 1780
 1781
 1782
 1783
 1784
 1785
 1786
 1787
 1788
 1789
 1790
 1791
 1792
 1793
 1794
- [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>. 1779
 1780
 1781
 1782
 1783
 1784
 1785
 1786
 1787
 1788
 1789
 1790
 1791
 1792
 1793
 1794
- [14] World Health Organization. Verbal Autopsy Standards: Ascertaining and Attributing Cause of Death. World Health Organization; 2007. 1787
 1788
 1789
 1790
 1791
 1792
 1793
 1794
- [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 1791
 1792
 1793
 1794

- 1795 Million Death Study. *Health Affairs*. 2017 Nov;36(11):1887–1895. <https://doi.org/10.1377/hlthaff.2017.0635>.
- 1796
- 1797
- 1798
- 1799 [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
- 1800 Results. *PLOS Medicine*. 2005 Dec;3(2):e18. <https://doi.org/10.1371/journal.pmed.0030018>.
- 1801
- 1802
- 1803
- 1804
- 1805
- 1806 [17] McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark SJ. Probabilistic Cause-of-death Assignment Using Verbal Autopsies. *Journal of the*
- 1807 American Statistical Association. 2016;111(515):1036–1049. <https://doi.org/10.1080/01621459.2016.1152191>.
- 1808
- 1809
- 1810
- 1811
- 1812
- 1813
- 1814 [18] Morris SK, Bassani DG, Kumar R, Awasthi S, Paul VK, Jha P. Factors Associated
- 1815 with Physician Agreement on Verbal Autopsy of over 27000 Childhood Deaths in
- 1816 India. *PLoS one*. 2010;5(3):e9583.
- 1817
- 1818
- 1819
- 1820 [19] Soleman N, Chandramohan D, Shibuya K. Verbal Autopsy: Current Practices
- 1821 and Challenges. *Bulletin of the World Health Organization*. 2006;84(3):239–245.
- 1822
- 1823
- 1824 [20] Byass P, Hussain-Alkhateeb L, D'Ambruoso L, Clark S, Davies J, Fottrell E,
- 1825 et al. An Integrated Approach to Processing WHO-2016 Verbal Autopsy Data:
- 1826 The InterVA-5 Model. *BMC Medicine*. 2019 May;17(1):102. <https://doi.org/10.1186/s12916-019-1333-6>.
- 1827
- 1828
- 1829
- 1830
- 1831 [21] Jha P, Kumar D, Dikshit R, Budukh A, Begum R, Sati P, et al. Automated versus
- 1832 Physician Assignment of Cause of Death for Verbal Autopsies: Randomized Trial
- 1833 of 9374 Deaths in 117 Villages in India. *BMC Medicine*. 2019 Jun;17(1):116.
- 1834
- 1835
- 1836
- 1837
- 1838
- 1839
- 1840

- [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>. 1841
1842
1843
1844
1845
1846
1847
1848
1849
- [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>. 1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
- [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>. 1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
- [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>. 1878
1879
1880
1881
1882
1883
1884
1885
1886
- [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>. 1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
- [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>. 1897
1898
1899
1900
1901
1902
1903
1904
1905
1906

- 1887 [28] King C, Zamawe C, Banda M, Bar-Zeev N, Beard J, Bird J, et al. The Quality and
1888 Diagnostic Value of Open Narratives in Verbal Autopsy: A Mixed-Methods Anal-
1889 ysis of Partnered Interviews from Malawi. BMC Medical Research Methodology.
1890 2016 Feb;16(1):13. <https://doi.org/10.1186/s12874-016-0115-5>.
1891
1892
1893
1894 [29] Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al.: A Survey on Evaluation
1895 of Large Language Models. arXiv.
1896
1897
1898 [30] Lund BD, Wang T. Chatting about ChatGPT: How May AI and GPT Impact
1899 Academia and Libraries? Library Hi Tech News. 2023 Jan;40(3):26–29. <https://doi.org/10.1108/LHTN-01-2023-0009>.
1900
1901
1902
1903
1904 [31] Svyatkovskiy A, Deng SK, Fu S, Sundaresan N. IntelliCode Compose: Code
1905 Generation Using Transformer. In: Proceedings of the 28th ACM Joint Meeting on
1906 European Software Engineering Conference and Symposium on the Foundations
1907 of Software Engineering. ESEC/FSE 2020. New York, NY, USA: Association for
1908 Computing Machinery; 2020. p. 1433–1443.
1909
1910
1911
1912
1913 [32] Haupt CE, Marks M. AI-Generated Medical Advice—GPT and Beyond. JAMA.
1914 2023 Apr;329(16):1349–1350. <https://doi.org/10.1001/jama.2023.5321>.
1915
1916
1917
1918 [33] Wu T, He S, Liu J, Sun S, Liu K, Han QL, et al. A Brief Overview of ChatGPT:
1919 The History, Status Quo and Potential Future Development. IEEE/CAA Journal
1920 of Automatica Sinica. 2023;10(5):1122–1136. <https://doi.org/10.1109/JAS.2023.123618>.
1921
1922
1923
1924
1925 [34] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al.: GPT-4
1926 Technical Report. arXiv.
1927
1928
1929 [35] Njala University.: Healthy Sierra Leone. <https://healsl.org/>.
1930
1931
1932

- [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](https://doi.org/10.1016/S2214-109X(21)00459-9).
- [37] World Health Organization. ICD-10: International Statistical Classification of Diseases and Related Health Problems (10th Revision); 2011.
- [38] Aleksandrowicz L, Malhotra V, Dikshit R, Gupta PC, Kumar R, Sheth J, et al. Performance Criteria for Verbal Autopsy-Based Systems to Estimate National Causes of Death: Development and Application to the Indian Million Death Study. *BMC Medicine*. 2014 Feb;12(1):21. <https://doi.org/10.1186/1741-7015-12-21>.
- [39] Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. *JAMA Network Open*. 2019 Mar;2(3):e190096. <https://doi.org/10.1001/jamanetworkopen.2019.0096>.
- [40] Hsiao M, Morris SK, Bassani DG, Montgomery AL, Thakur JS, Jha P. Factors Associated with Physician Agreement on Verbal Autopsy of over 11500 Injury Deaths in India. *PLOS ONE*. 2012 Jan;7(1):e30336. <https://doi.org/10.1371/journal.pone.0030336>.
- [41] Murray CJ, Lozano R, Flaxman AD, Serina P, Phillips D, Stewart A, et al. Using Verbal Autopsy to Measure Causes of Death: The Comparative Performance of Existing Methods. *BMC Medicine*. 2014 Jan;12(1):5. <https://doi.org/10.1186/1741-7015-12-5>.

- 1979 [42] Benara SK, Sharma S, Juneja A, Nair S, Gulati BK, Singh KJ, et al. Evaluation of
1980 Methods for Assigning Causes of Death from Verbal Autopsies in India. *Frontiers*
1981 in Big Data.
- 1982 2023 Aug;6:1197471. <https://doi.org/10.3389/fdata.2023.1197471>.
- 1983
- 1984
- 1985 [43] Li ZR, Thomas J, Choi E, McCormick TH, Clark SJ. The openVA Toolkit for
1986 Verbal Autopsies. *The R Journal*. 2023 Feb;p. 1.
- 1987
- 1988
- 1989 [44] BAYES. An Essay towards Solving a Problem in the Doctrine of Chances.
1990 Biometrika. 1958;45(3-4):296–315.
- 1991
- 1992
- 1993 [45] Byass P, Chandramohan D, Clark SJ, D'Ambruoso L, Fottrell E, Graham WJ,
1994 et al. Strengthening Standardised Interpretation of Verbal Autopsy Data: The
1995 New InterVA-4 Tool. *Global Health Action*. 2012 Dec;5(1):19281. <https://doi.org/10.3402/gha.v5i0.19281>.
- 1996
- 1997
- 1998
- 1999
- 2000
- 2001 [46] Brooks S. Markov Chain Monte Carlo Method and Its Application. *Journal of*
2002 the Royal Statistical Society: Series D (The Statistician).
- 2003 1998 Mar;47(1):69–100.
- 2004 <https://doi.org/10.1111/1467-9884.00117>.
- 2005
- 2006
- 2007 [47] Chib S. Markov Chain Monte Carlo Methods: Computation and Inference.
- 2008 Handbook of econometrics.
- 2009 2001;5:3569–3649.
- 2010
- 2011 [48] Han C, Carlin BP. Markov Chain Monte Carlo Methods for Computing Bayes
2012 Factors: A Comparative Review. *Journal of the American Statistical Association*.
- 2013
- 2014 2001 Sep;96(455):1122–1132. <https://doi.org/10.1198/016214501753208780>.
- 2015
- 2016
- 2017 [49] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al.: Language
2018 Models Are Few-Shot Learners. arXiv.
- 2019
- 2020
- 2021 [50] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Atten-
2022 tion Is All You Need. In: Advances in Neural Information Processing Systems.
- 2023
- 2024

- vol. 30. Curran Associates, Inc.; 2017. . 2025
- [51] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al.: Training 2026
Language Models to Follow Instructions with Human Feedback. arXiv. 2027
2028
- [52] Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep Rein- 2029
forcement Learning from Human Preferences. Advances in neural information 2030
processing systems. 2017;30. 2031
2032
- [53] Stiennon N, Ouyang L, Wu J, Ziegler D, Lowe R, Voss C, et al. Learning to 2033
Summarize with Human Feedback. Advances in Neural Information Processing 2034
Systems. 2020;33:3008–3021. 2035
2036
- [54] Wirth C, Akrou R, Neumann G, Fürnkranz J. A Survey of Preference-Based 2037
Reinforcement Learning Methods. The Journal of Machine Learning Research. 2038
2017 Jan;18(1):4945–4990. 2039
2040
- [55] World Health Organization.: Verbal Autopsy Standards: The 2016 WHO Ver- 2041
bal Autopsy Instrument. <https://www.who.int/publications/m/item/verbal-autopsy-standards-the-2016-who-verbal-autopsy-instrument>. 2042
- [56] Murray CJ, Lozano R, Flaxman AD, Vahdatpour A, Lopez AD. Robust Metrics 2043
for Assessing the Performance of Different Verbal Autopsy Cause Assignment 2044
Methods in Validation Studies. Population Health Metrics. 2011 Aug;9(1):28. 2045
<https://doi.org/10.1186/1478-7954-9-28>. 2046
2047
- [57] Setel PW, Whiting DR, Hemed Y, Chandramohan D, Wolfson LJ, Alberti 2048
KGMM, et al. Validity of Verbal Autopsy Procedures for Determining Cause of 2049
Death in Tanzania. Tropical Medicine & International Health. 2006;11(5):681– 2050
696. <https://doi.org/10.1111/j.1365-3156.2006.01603.x>. 2051
2052
- 2053
2054
- 2055
2056
- 2057
2058
- 2059
2060
- 2061
2062
- 2063
2064
- 2065
2066
- 2067
2068
- 2069
2070

- 2071 [58] World Health Organization. World Health Statistics 2024: Monitoring Health for
2072 the SDGs, Sustainable Development Goals. World Health Organization; 2024.
2073
- 2074
- 2075 [59] World Health Organization.: Cause-Specific Mortality 2000 and 2021. The Global
2076 Health Observatory.
2077
- 2078
- 2079 [60] Rasmussen LA, Cascio MA, Ferrand A, Shevell M, Racine E. The Complex-
2080 ity of Physicians' Understanding and Management of Prognostic Uncertainty
2081 in Neonatal Hypoxic-Ischemic Encephalopathy. Journal of Perinatology. 2019
2082 Feb;39(2):278–285. <https://doi.org/10.1038/s41372-018-0296-3>.
2083
- 2084
- 2085
- 2086
- 2087 [61] Faison G, Chou FS, Feudtner C, Janvier A. When the Unknown Is Unknowable:
2088 Confronting Diagnostic Uncertainty. Pediatrics. 2023 Sep;152(4):e2023061193.
2089 <https://doi.org/10.1542/peds.2023-061193>.
2090
- 2091
- 2092
- 2093 [62] OpenAI.: Pricing. <https://openai.com/api/pricing/>.
2094
- 2095 [63] Tao G, Cheng S, Zhang Z, Zhu J, Shen G, Zhang X.: Opening A Pandora's Box:
2096 Things You Should Know in the Era of Custom GPTs. arXiv.
2097
- 2098
- 2099 [64] Khowaja SA, Khuwaja P, Dev K, Wang W, Nkenyereye L. ChatGPT
2100 Needs SPADE (Sustainability, PrivAcy, Digital Divide, and Ethics) Evalu-
2101 ation: A Review. Cognitive Computation. 2024 May;<https://doi.org/10.1007/s12559-024-10285-1>.
2102
- 2103
- 2104
- 2105
- 2106
- 2107 [65] Wu X, Duan R, Ni J. Unveiling Security, Privacy, and Ethical Concerns of
2108 ChatGPT. Journal of Information and Intelligence. 2024;2(2):102–115.
2109
- 2110
- 2111 [66] OpenAI.: OpenAI Platform: API Reference (Tempera-
2112 ture Parameter). [https://platform.openai.com/docs/api-](https://platform.openai.com/docs/api-reference/completions/create#completions-create-temperature)
2113 reference/completions/create#completions-create-temperature.
2114
- 2115
- 2116

- [67] Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Research Square. 2023 Feb;p. rs.3.rs-2566942. <https://doi.org/10.21203/rs.3.rs-2566942/v1>. 2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
- [68] Jang ME, Lukasiewicz T.: Consistency Analysis of ChatGPT. arXiv. 2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
- [69] Krishna S, Bhambra N, Bleakney R, Bhayana R, Atzen S. Evaluation of Reliability, Repeatability, Robustness, and Confidence of GPT-3.5 and GPT-4 on a Radiology Board-Style Examination. Radiology. 2024 May;311(2):e232715. <https://doi.org/10.1148/radiol.232715>. 2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
- [70] Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, et al.: Prompt Engineering for Healthcare: Methodologies and Applications. arXiv. 2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
- [71] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20. Red Hook, NY, USA: Curran Associates Inc.; 2020. p. 9459–9474. 2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
- [72] Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. Journal of medical Internet research. 2023;25:e50638. 2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
- [73] Almasre M. Development and Evaluation of a Custom GPT for the Assessment of Students' Designs in a Typography Course. Education Sciences. 2024 Feb;14(2):148. <https://doi.org/10.3390/educsci14020148>. 2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
- [74] Loh P, Fottrell E, Beard J, Bar-Zeev N, Phiri T, Banda M, et al. Added Value of an Open Narrative in Verbal Autopsies: A Mixed-Methods Evaluation from Malawi. BMJ Paediatrics Open. 2021 Feb;5(1):e000961. <https://doi.org/10.1136/bmjpo-2020-000961>. 2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162

2163 [75] World Health Organization.: ODK for Verbal Autopsy: A Quick Guide.
2164
2165 https://www.who.int/publications/m/item/odk-for-verbal-autopsy-a-quick-
2166 guide.
2167
2168
2169 [76] Nafundi.: ODK - Collect Data Anywhere.
2170
2171 [77] DiPasquale A, Maire N, Bratschi M.: Release ODK 2016 WHO VA Instrument
2172
2173 1.5.1 SwissTPH/WHO-VA. Swiss Tropical and Public Health Institute.
2174
2175 [78] Byass P.: InterVA-5.1 User Guide.
2176
2177
2178 [79] Thomas J, ekarpinskiMITRE, pkmitre, owentrigueros, Choi P, Chu Y.: Pycrossva:
2179 Prepare Data from WHO and PHRMC Instruments for Verbal Autopsy Algo-
2180
2181 rithms.
2182
2183
2184 [80] Li ZR, McCormick T, Clark S.: InSilicoVA: Probabilistic Verbal Autopsy Coding
2185 with 'InSilicoVA' Algorithm.
2186
2187
2188 [81] Thomas J, Li Z, Byass P, McCormick T, Boyas M, Clark S.: InterVA5: Replicate
2189 and Analyse 'InterVA5'.
2190
2191
2192 [82] Yendewa GA, Poveda E, Yendewa SA, Sahr F, Quiñones-Mateu ME, Salata RA.
2193
2194 HIV/AIDS in Sierra Leone: Characterizing the Hidden Epidemic. AIDS reviews.
2195
2196 2018;20(2).
2197
2198 [83] Walker PG, White MT, Griffin JT, Reynolds A, Ferguson NM, Ghani AC. Malaria
2199 Morbidity and Mortality in Ebola-affected Countries Caused by Decreased
2200
2201 Health-Care Capacity, and the Potential Effect of Mitigation Strategies: A
2202
2203 Modelling Analysis. The Lancet Infectious Diseases. 2015;15(7):825–832.
2204
2205
2206
2207
2208