From Narratives to Numbers: Valid Inference Using Language Model Predictions from Verbal Autopsies

Adam Visokay

University of Washington

120TH ASA ANNUAL MEETING

Reimagining the Future of Work

ASA AUGUST 8-12, 2025
CHICAGO, IL

Experimental and Applied Social Research

Content warning

Discussing mortality.

Includes some textual descriptions of death/dying.

The Team

Adam Visokay

Trinity Fan

Kentaro Hoffman

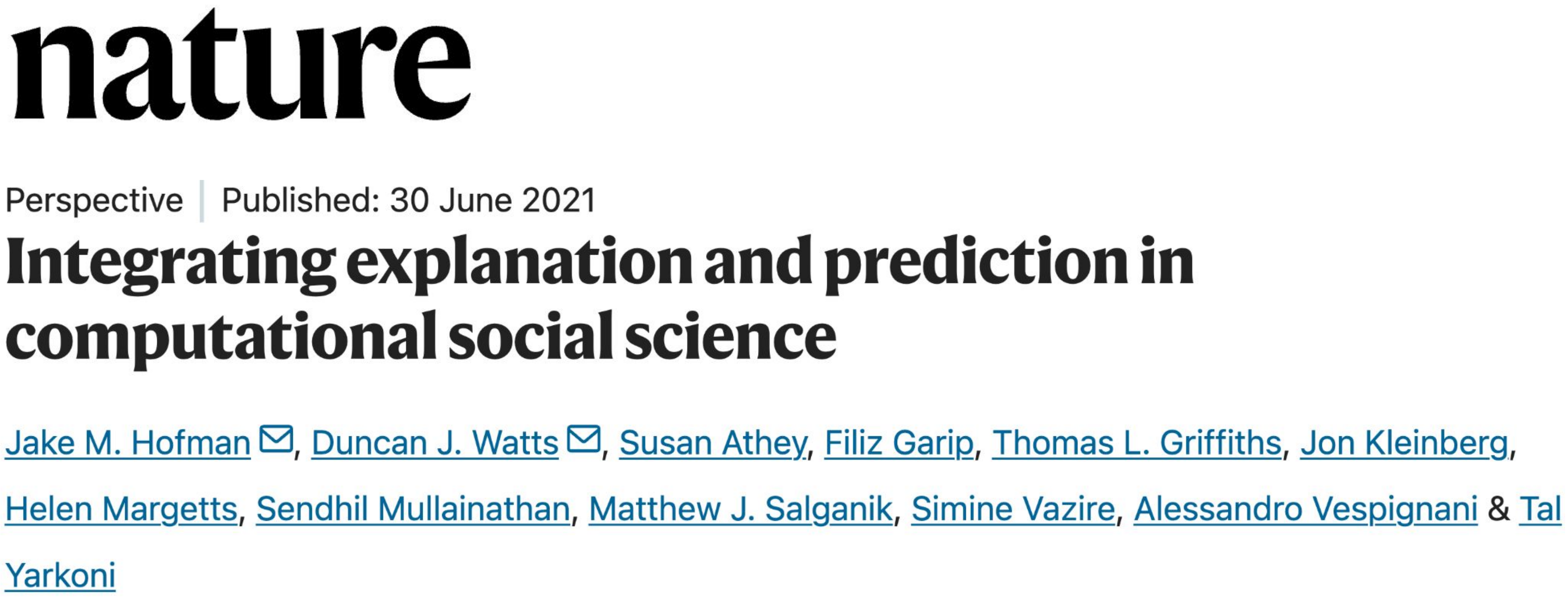
Stephen Salerno

Jeff Leek

Li Liu

Tyler McCormick

Link to Paper



nature

Perspective | Published: 30 June 2021

Integrating explanation and prediction in computational social science

Jake M. Hofman ✉, Duncan J. Watts ✉, Susan Athey, Filiz Garip, Thomas L. Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J. Salganik, Simine Vazire, Alessandro Vespignani & Tal Yarkoni

"Advocate for more *integrative modelling* …

… combining **prediction** and **explanation**".

$$y = \hat{\beta}\, X_1$$

$$y = \hat{\beta}\, X_1 \qquad \hat{y} = \hat{\beta}\, X_2$$

$$y = \hat{\beta} X_1$$

Prediction

$$\hat{y}$$

NLP

GPT-4

$$\hat{y} = \hat{\beta} X_2$$

Explanation

$$\hat{\beta}$$

Inference with Predicted Data

Domain

Low resource settings where most deaths are not reflected in official statistics (vital registration systems)

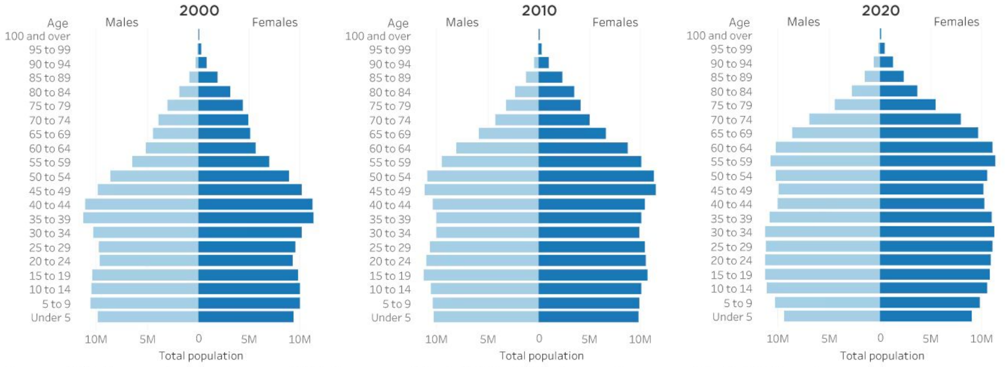
Question

How are changing population demographics (e.g. age) associated with changing cause of death distribution?

Figure 2.

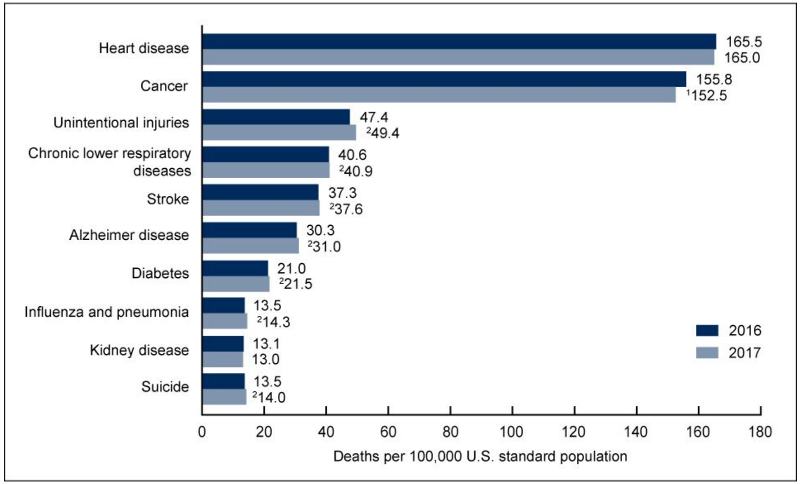
Population Pyramids for the United States: 2000, 2010 and 2020

(In millions)



Source: U.S. Census Bureau, Census 2000 Summary File 1 (SF1), 2010 Census Summary File 1 (SF1) and 2020 Census Demographic and Housing Characteristics File (DHC).

Figure 4. Age-adjusted death rates for the 10 leading causes of death: United States, 2016 and 2017



[1] Statistically significant decrease in age-adjusted death rate from 2016 to 2017 ($p < 0.05$).

[2] Statistically significant increase in age-adjusted death rate from 2016 to 2017 ($p < 0.05$).

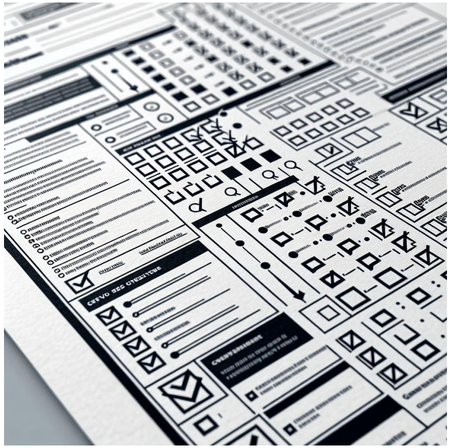
NOTES: A total of 2,813,503 resident deaths were registered in the United States in 2017. The 10 leading causes accounted for 74.0% of all deaths in the United States in 2017. Causes of death are ranked according to number of deaths. Rankings for 2016 data are not shown. Data table for Figure 4 includes the number of deaths for leading causes. Access data table for Figure 4 at: https://www.cdc.gov/nchs/data/databriefs/db328_tables-508.pdf#4.

SOURCE: NCHS, National Vital Statistics System, Mortality.

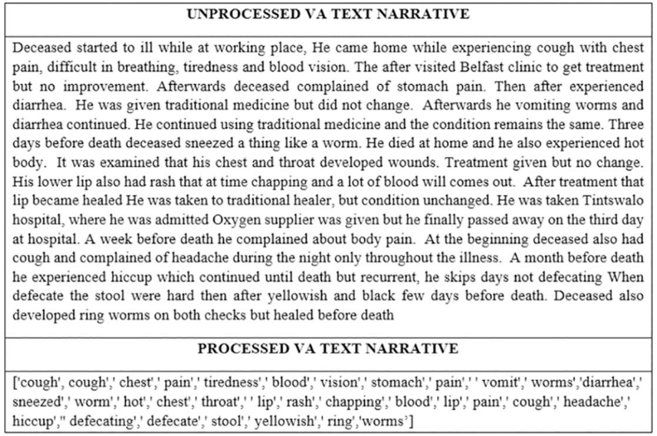
Verbal Autopsies

Interviews with caregivers of the deceased, used to assign COD.

structured questionnaire



free text narrative



UNPROCESSED VA TEXT NARRATIVE
Deceased started to ill while at working place, He came home while experiencing cough with chest pain, difficult in breathing, tiredness and blood vision. The after visited Belfast clinic to get treatment but no improvement. Afterwards deceased complained of stomach pain. Then after experienced diarrhea. He was given traditional medicine but did not change. Afterwards he vomiting worms and diarrhea continued. He continued using traditional medicine and the condition remains the same. Three days before death deceased sneezed a thing like a worm. He died at home and he also experienced hot body. It was examined that his chest and throat developed wounds. Treatment given but no change. His lower lip also had rash that at time chapping and a lot of blood will comes out. After treatment that lip became healed He was taken to traditional healer, but condition unchanged. He was taken Tintswalo hospital, where he was admitted Oxygen supplier was given but he finally passed away on the third day at hospital. A week before death he complained about body pain. At the beginning deceased also had cough and complained of headache during the night only throughout the illness. A month before death he experienced hiccup which continued until death but recurrent, he skips days not defecating When defecate the stool were hard then after yellowish and black few days before death. Deceased also developed ring worms on both checks but healed before death
PROCESSED VA TEXT NARRATIVE
['cough', cough',' chest',' pain',' tiredness',' blood',' vision',' stomach',' pain',' ' vomit',' worms','diarrhea',' sneezed',' worm',' hot',' chest',' throat',' ' lip',' rash',' chapping',' blood',' lip',' pain',' cough',' headache',' hiccup','' defecating',' defecate',' stool',' yellowish',' ring','worms']

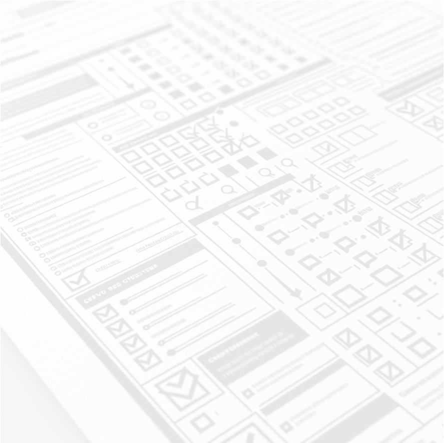
Mapundu et al. 2024

Burdensome on respondents (~2hr, repetitive, impersonal).

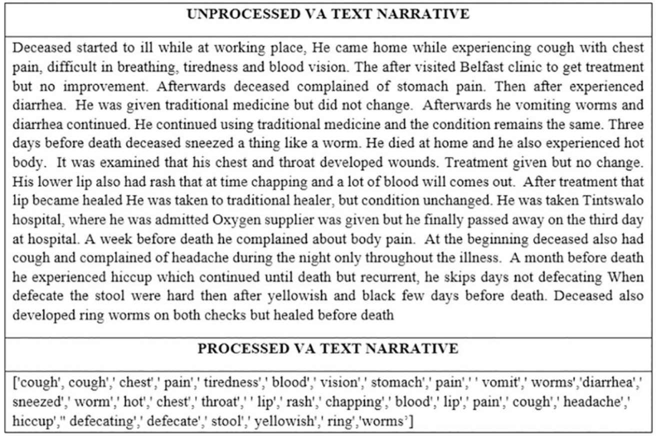
Verbal Autopsies

Interviews with caregivers of the deceased, used to assign COD.

structured questionnaire



free text narrative



UNPROCESSED VA TEXT NARRATIVE

Deceased started to ill while at working place, He came home while experiencing cough with chest pain, difficult in breathing, tiredness and blood vision. The after visited Belfast clinic to get treatment but no improvement. Afterwards deceased complained of stomach pain. Then after experienced diarrhea. He was given traditional medicine but did not change. Afterwards he vomiting worms and diarrhea continued. He continued using traditional medicine and the condition remains the same. Three days before death deceased sneezed a thing like a worm. He died at home and he also experienced hot body. It was examined that his chest and throat developed wounds. Treatment given but no change. His lower lip also had rash that at time chapping and a lot of blood will comes out. After treatment that lip became healed He was taken to traditional healer, but condition unchanged. He was taken Tintswalo hospital, where he was admitted Oxygen supplier was given but he finally passed away on the third day at hospital. A week before death he complained about body pain. At the beginning deceased also had cough and complained of headache during the night only throughout the illness. A month before death he experienced hiccup which continued until death but recurrent, he skips days not defecating When defecate the stool were hard then after yellowish and black few days before death. Deceased also developed ring worms on both checks but healed before death

PROCESSED VA TEXT NARRATIVE

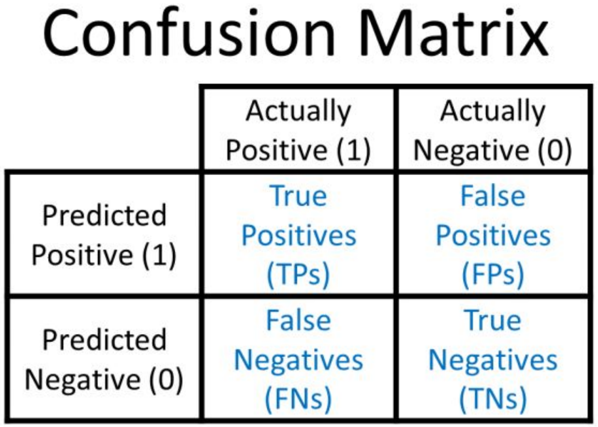
['cough', cough',' chest',' pain',' tiredness',' blood',' vision',' stomach',' pain',' ' vomit',' worms','diarrhea',' sneezed',' worm',' hot',' chest',' throat',' ' lip',' rash',' chapping',' blood',' lip',' pain',' cough',' headache',' hiccup','' defecating',' defecate',' stool',' yellowish',' ring','worms']

Mapundu et al. 2024

Burdensome on respondents (~2hr, repetitive, impersonal).

Motivation

You use an AI/ML algorithm to make predictions.



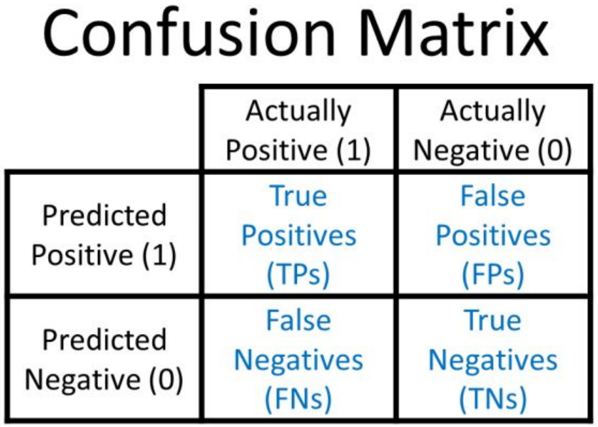
Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Motivation

You use an AI/ML algorithm to make predictions. Now what?



Confusion Matrix

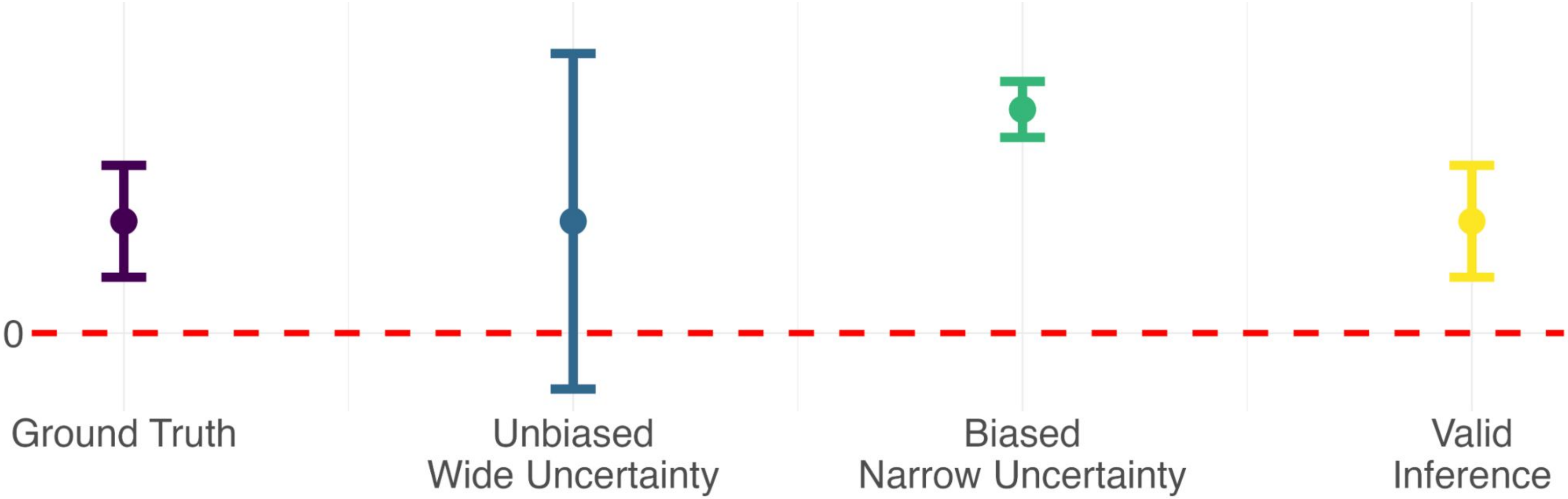
	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

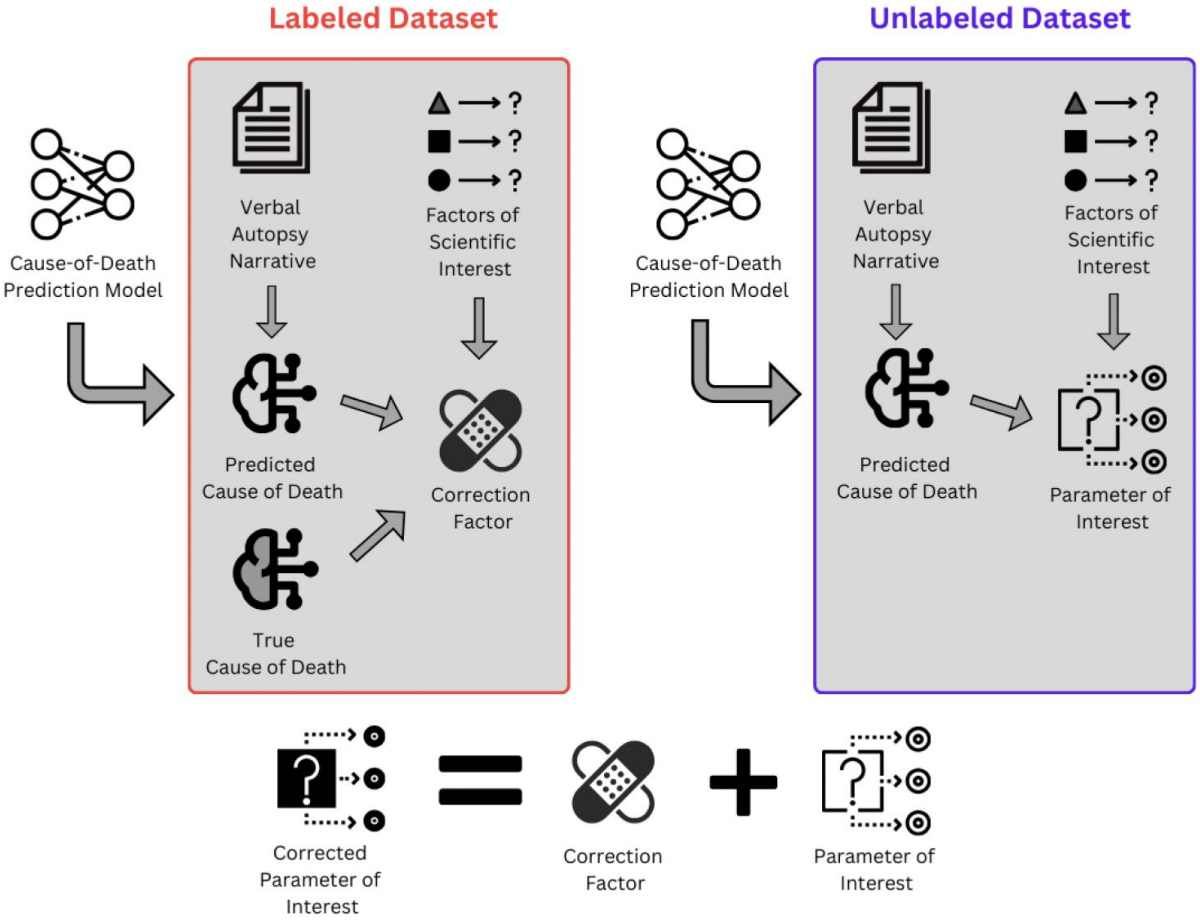
?

Inference with predicted data (IPD) can have:

1. Biased estimates
2. Misleading uncertainty



Inference with Predicted Data (multiPPI++)



Labeled Dataset

Cause-of-Death Prediction Model

Verbal Autopsy Narrative

▲ → ?
■ → ?
● → ?

Factors of Scientific Interest

Predicted Cause of Death

Correction Factor

True Cause of Death

Unlabeled Dataset

Cause-of-Death Prediction Model

Verbal Autopsy Narrative

▲ → ?
■ → ?
● → ?

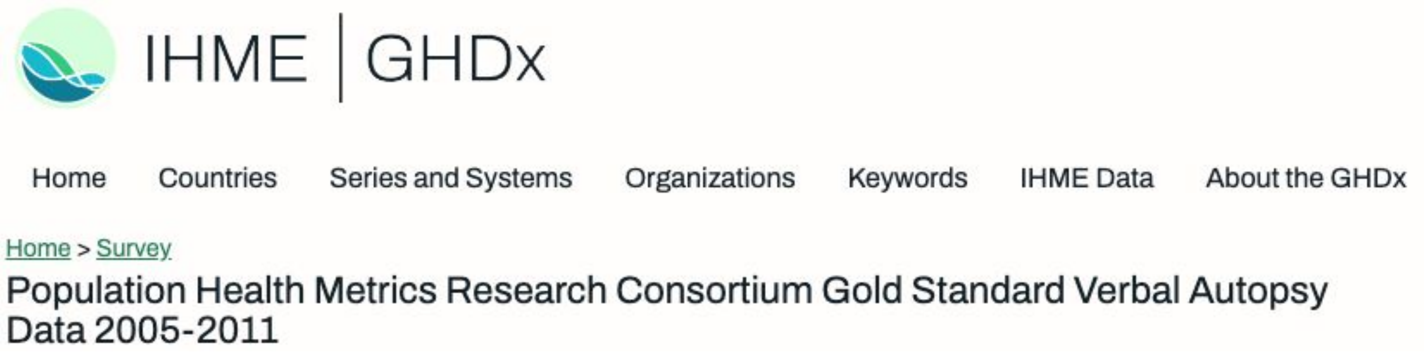
Factors of Scientific Interest

Predicted Cause of Death

Parameter of Interest

Corrected Parameter of Interest = Correction Factor + Parameter of Interest

Data



IHME | GHDx

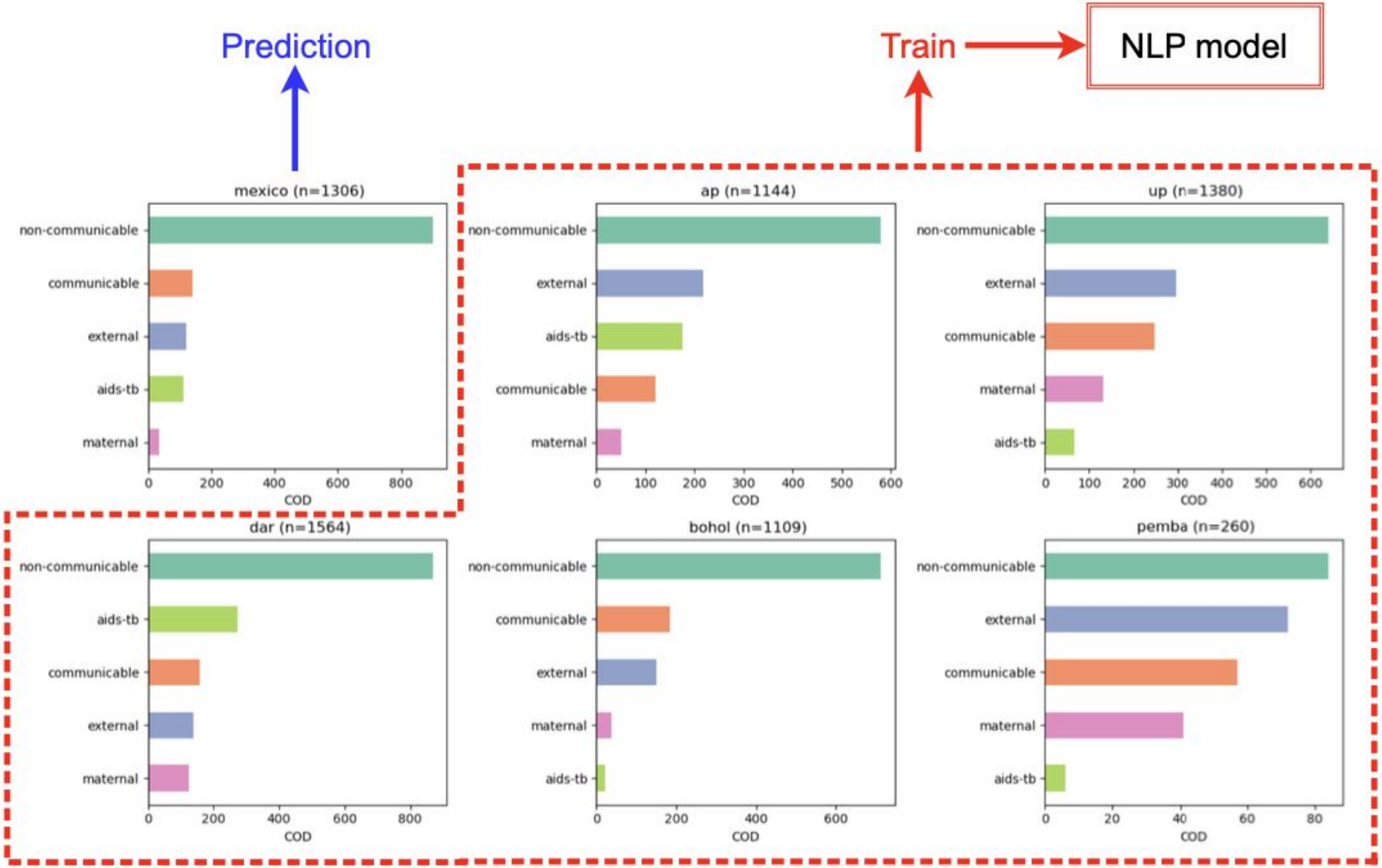
Population Health Metrics Research Consortium Gold Standard Verbal Autopsy Data 2005-2011

- adult deaths (n=6763)
- both traditional **and** verbal autopsies
- 6 sites, 4 countries
- 5 COD - [*Communicable, Non-communicable, Maternal, AIDS-TB, External*]

Validation set allows us to evaluate our experiment!

Experimental Design - leave one out validation

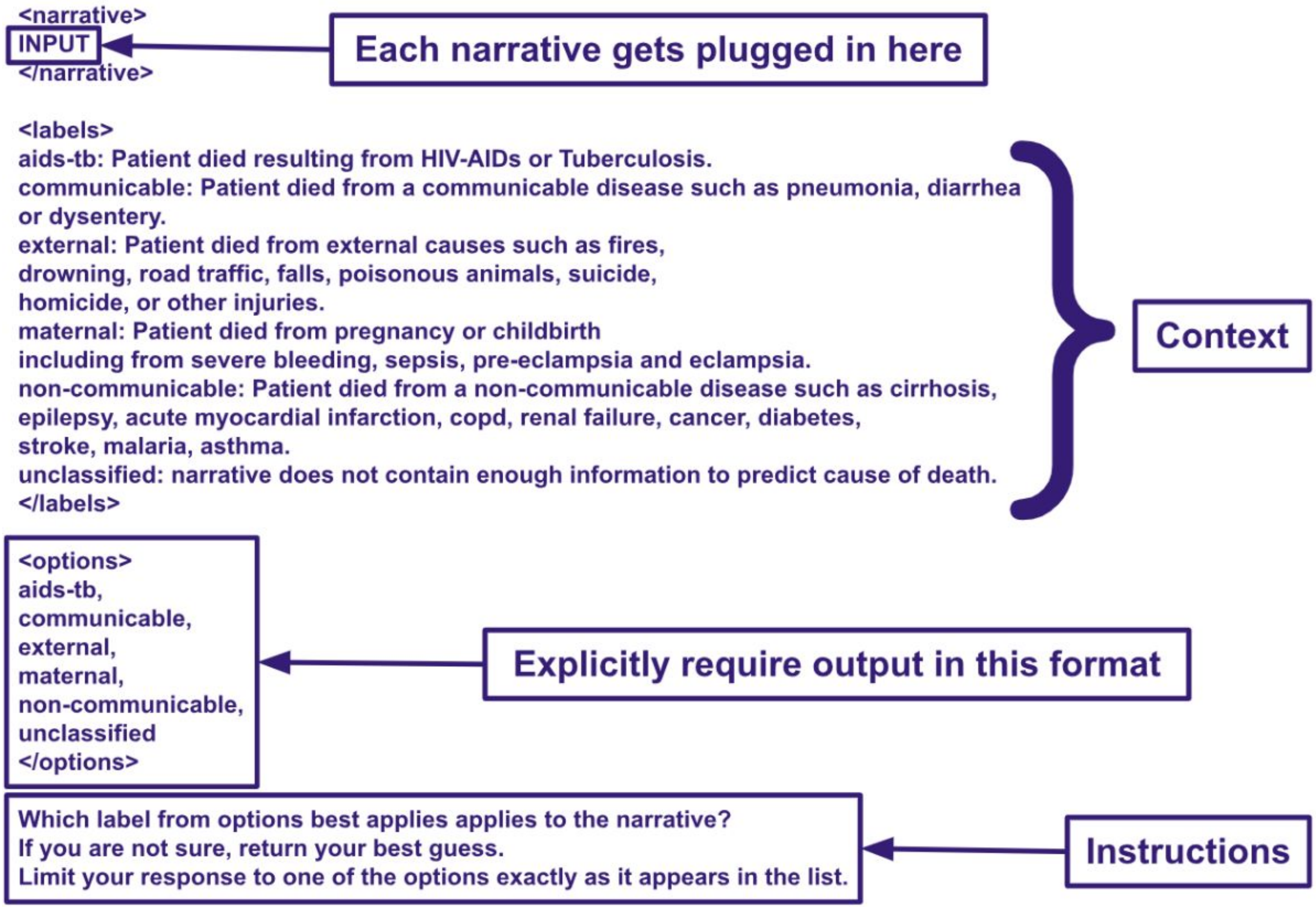
Bag of words (Naive Bayes, KNN, SVM), BERT, GPT-4



Prediction

Train

NLP model



```
<narrative>
INPUT
</narrative>

<labels>
aids-tb: Patient died resulting from HIV-AIDs or Tuberculosis.
communicable: Patient died from a communicable disease such as pneumonia, diarrhea
or dysentery.
external: Patient died from external causes such as fires,
drowning, road traffic, falls, poisonous animals, suicide,
homicide, or other injuries.
maternal: Patient died from pregnancy or childbirth
including from severe bleeding, sepsis, pre-eclampsia and eclampsia.
non-communicable: Patient died from a non-communicable disease such as cirrhosis,
epilepsy, acute myocardial infarction, copd, renal failure, cancer, diabetes,
stroke, malaria, asthma.
unclassified: narrative does not contain enough information to predict cause of death.
</labels>

<options>
aids-tb,
communicable,
external,
maternal,
non-communicable,
unclassified
</options>

Which label from options best applies applies to the narrative?
If you are not sure, return your best guess.
Limit your response to one of the options exactly as it appears in the list.
```

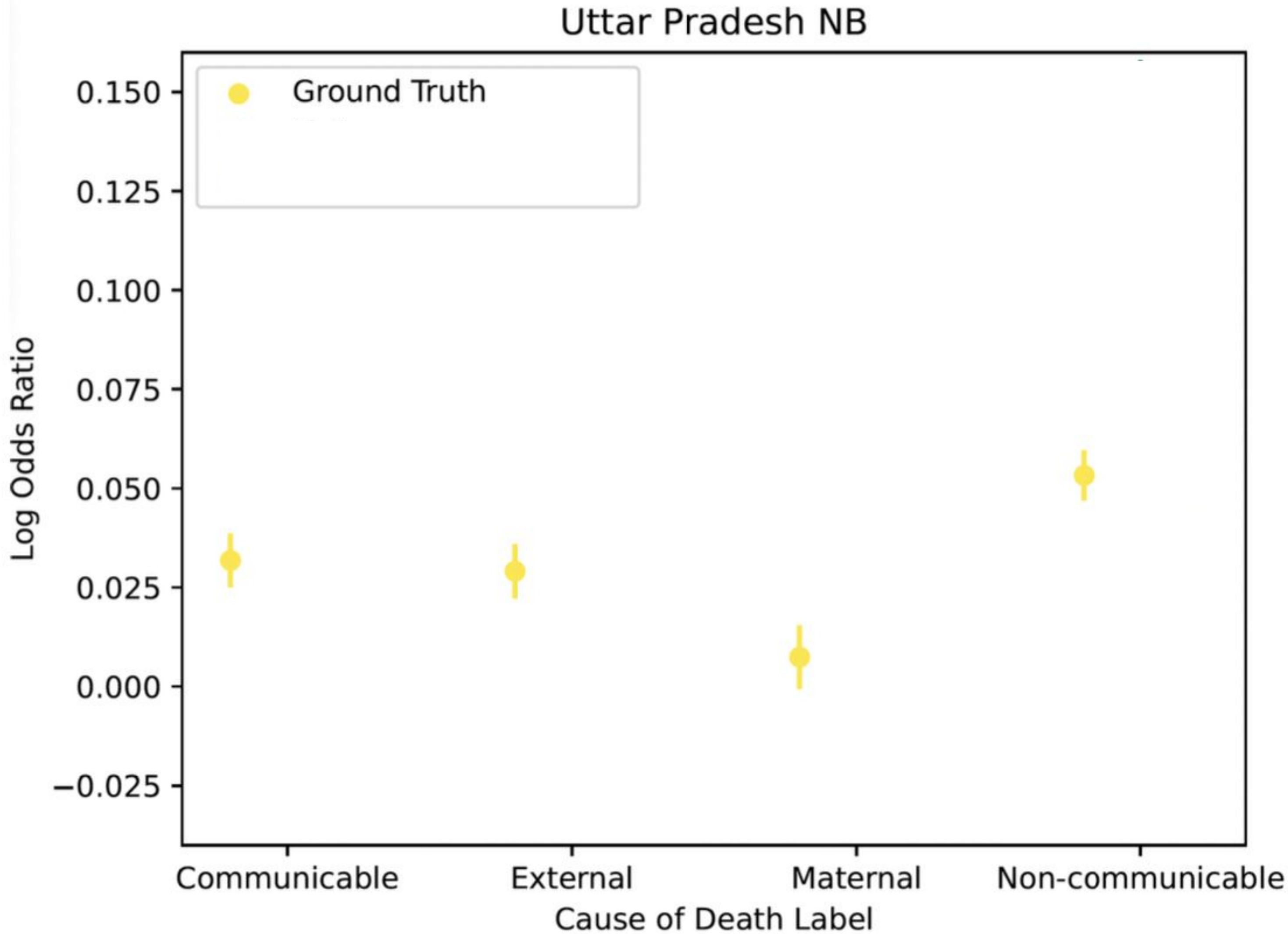
Each narrative gets plugged in here
Context
Explicitly require output in this format
Instructions

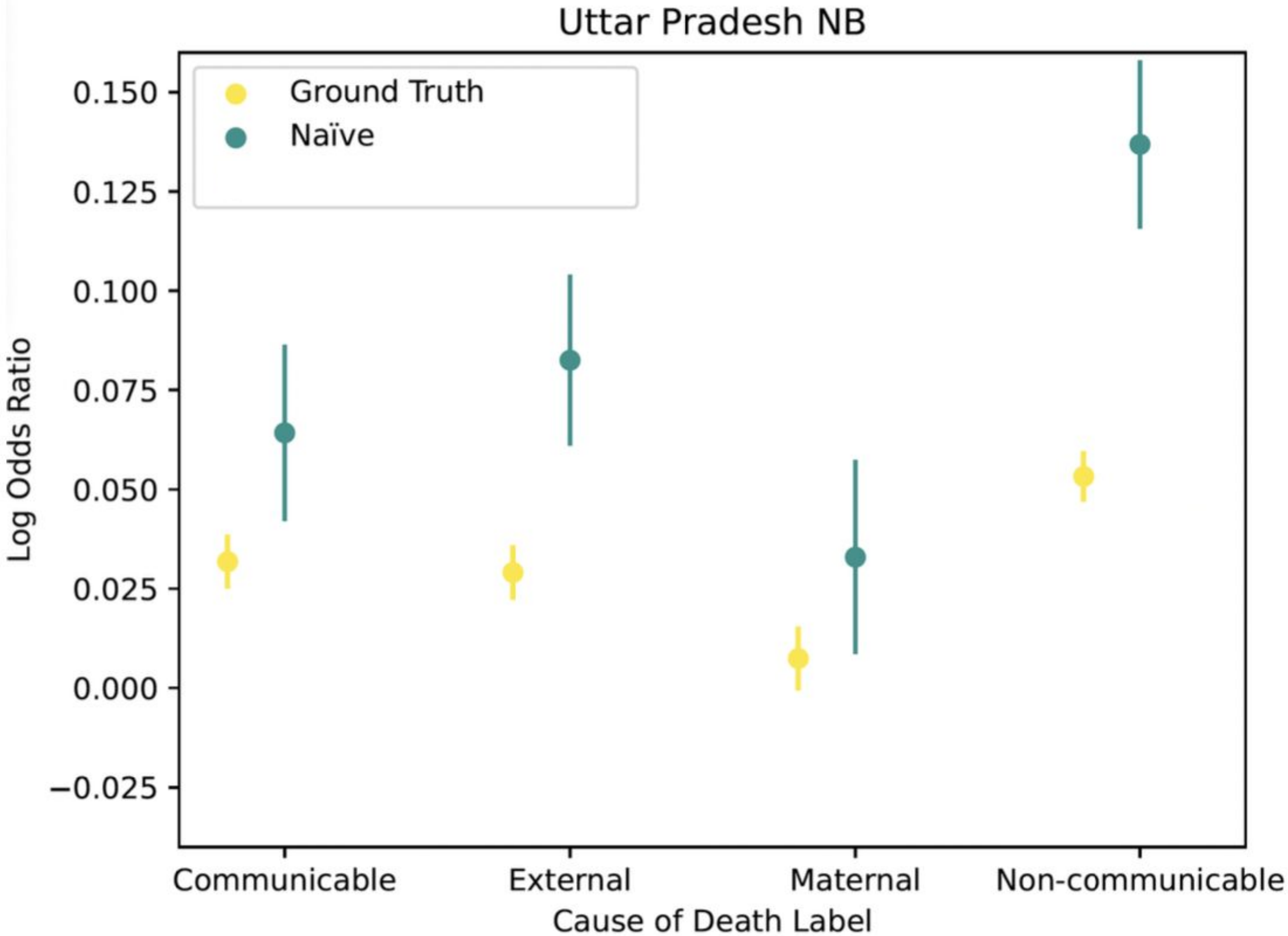
Multinomial Logistic Regression

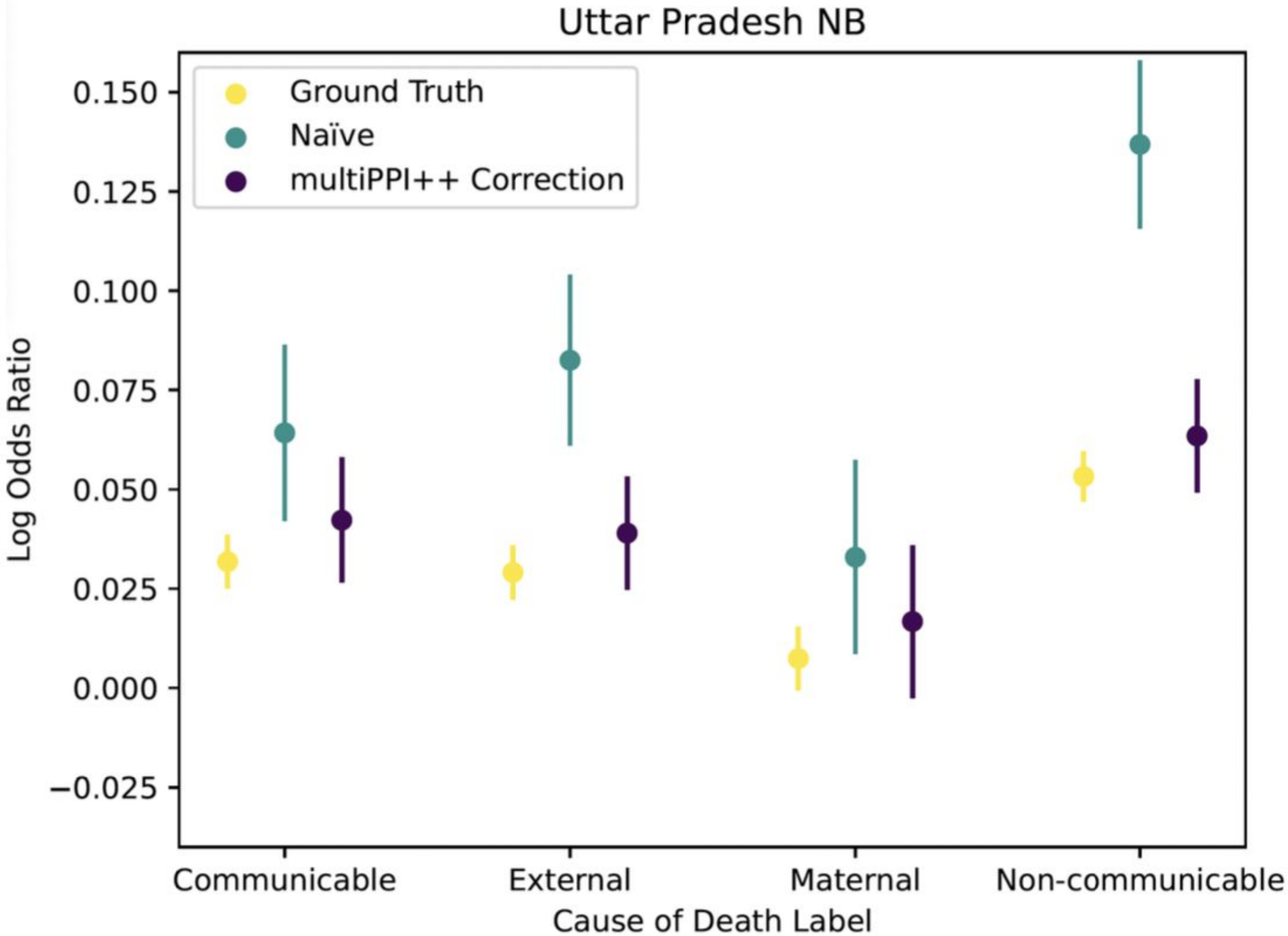
Cause specific mortality associated with Age.

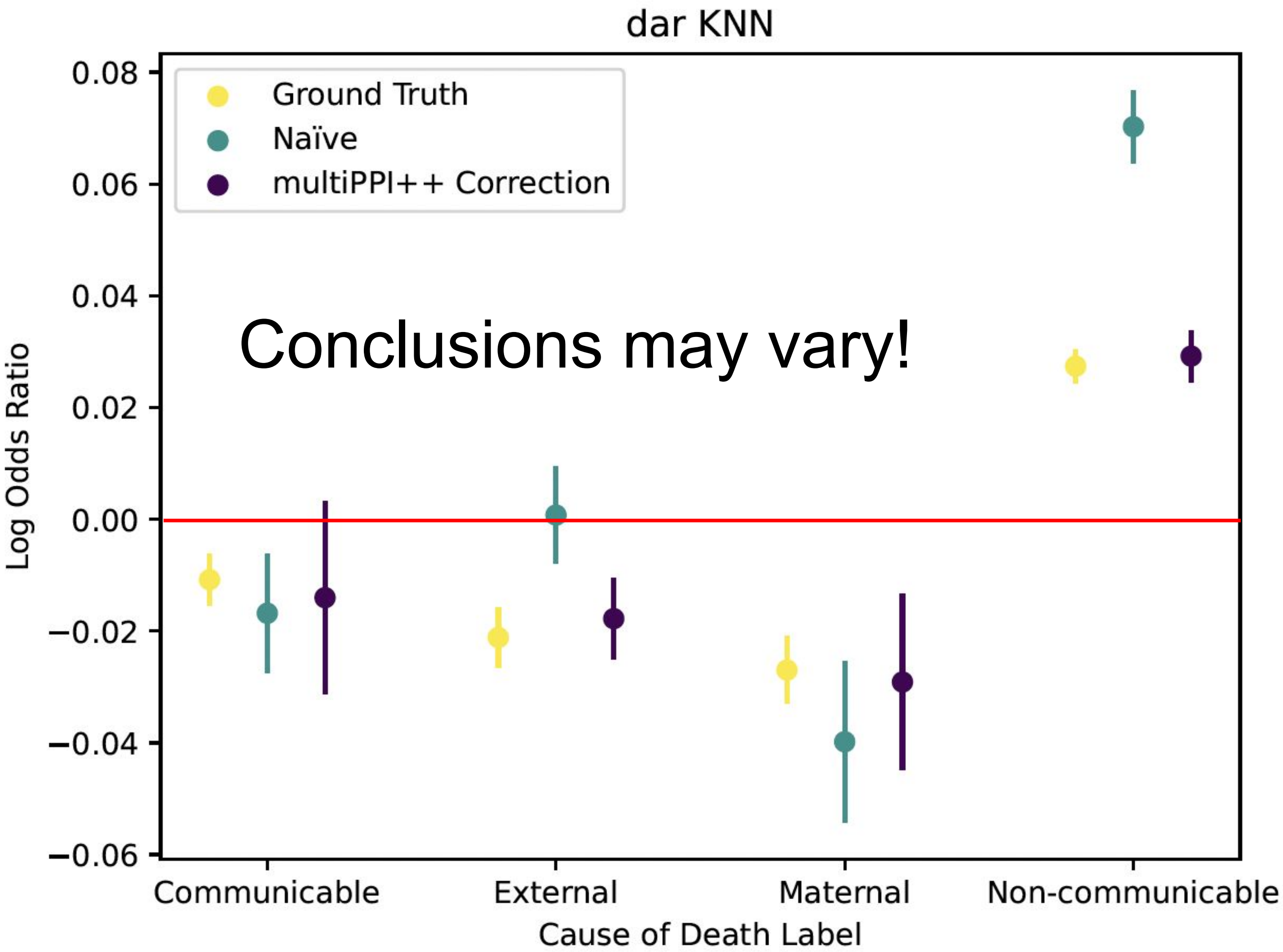
$$\log(\frac{p_{COD_i}}{p_{COD_{reference}}}) = \theta_0 + X_{age} * \theta_i$$

where θ_i is the change in log-odds of dying to cause *i* relative to the reference COD (aids-tb).


Uttar Pradesh NB
● Ground Truth
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
Log Odds Ratio
Communicable
External
Maternal
Non-communicable
Cause of Death Label


Uttar Pradesh NB
Ground Truth
Naïve
Log Odds Ratio
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
Communicable
External
Maternal
Non-communicable
Cause of Death Label


Uttar Pradesh NB
Ground Truth
Naïve
multiPPI++ Correction
Log Odds Ratio
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
Communicable
External
Maternal
Non-communicable
Cause of Death Label



Conclusions may vary!

Takeaways

1. Even with cutting edge models, data matters a lot!

Takeaways

1. Even with cutting edge models, data matters a lot!

2. Structured VA interviews are extremely burdensome (2-3 hours, redundant, impersonal).

Takeaways

1. Even with cutting edge models, data matters a lot!

2. Structured VA interviews are extremely burdensome (2-3 hours, redundant, impersonal).

   Narratives can be collected in 20-30 minutes …

   … and can actually be cathartic for respondents.

Lowes & Gill 2006, Connolly et al. 2023

Thank you!!

Contact:
Adam Visokay
avisokay@uw.edu
https://avisokay.github.io/

arxiv

IPD software is available!
Paper
Github
CRAN

Appendix

Regularized Loss Function

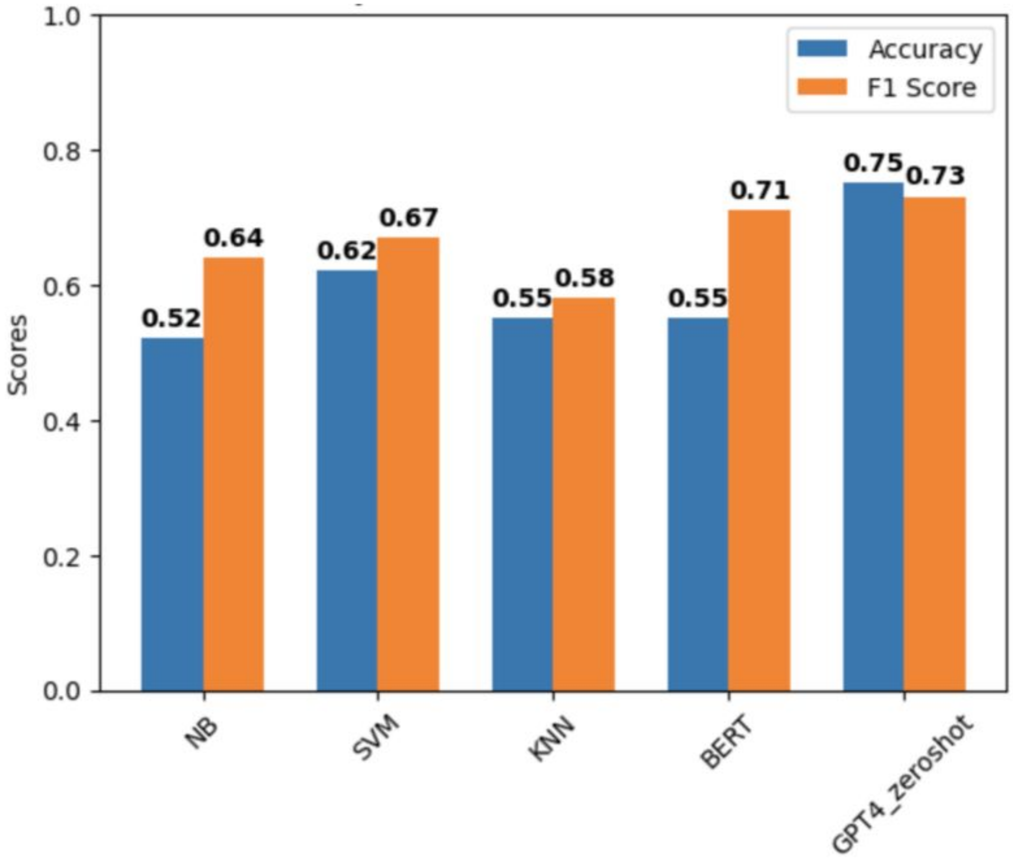
$$\mathbb{E}[\ell_\theta(X_L, Y_L)] \quad +$$

$$\lambda \left(\mathbb{E}[\ell_\theta(X_U, \hat{Y}_U^{AI})] - \mathbb{E}[l_\theta(X_L, \hat{Y}_L^{AI})] \right)$$

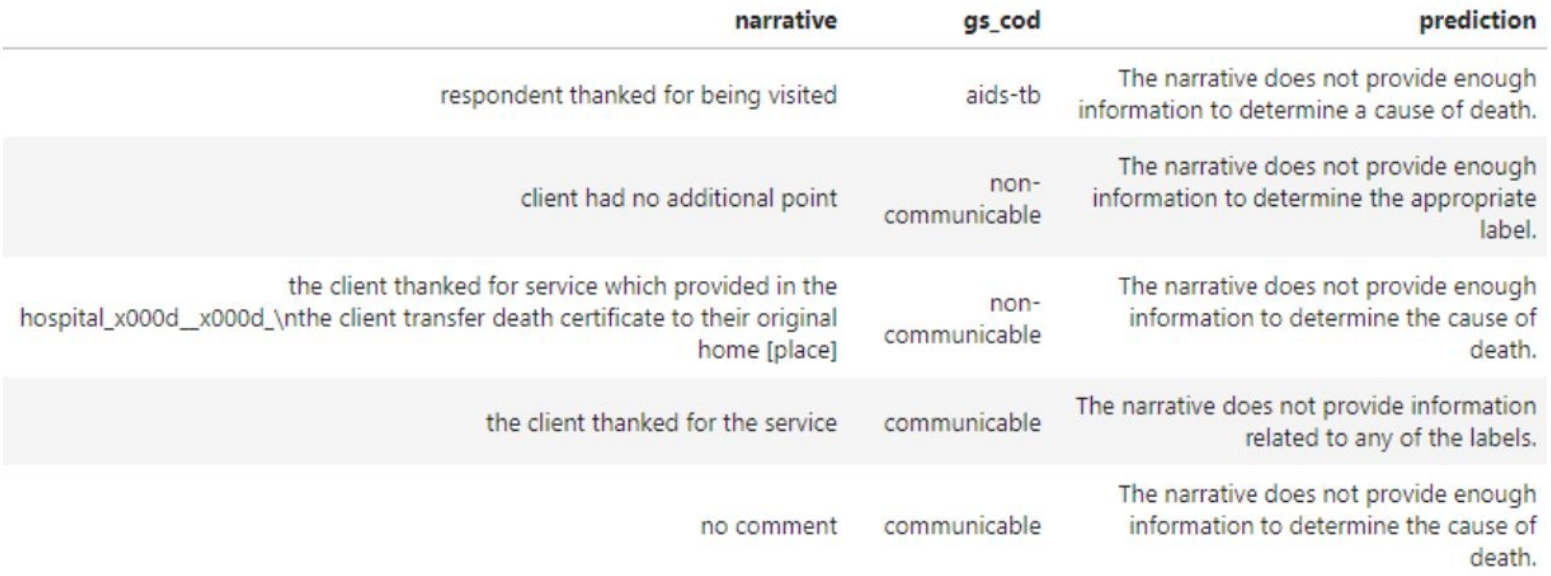
Lambda is a tuning parameter in [0,1]

Lambda = 0 when the predicted data are all **noise**

Lambda = 1 when the predicted data are all **signal**


1.0
0.8
0.6
0.4
0.2
0.0
Scores
Accuracy
F1 Score
0.52
0.64
0.62
0.67
0.55
0.58
0.55
0.71
0.75
0.73
NB
SVM
KNN
BERT
GPT4_zeroshot

Closer Look at GPT-4 Predictions

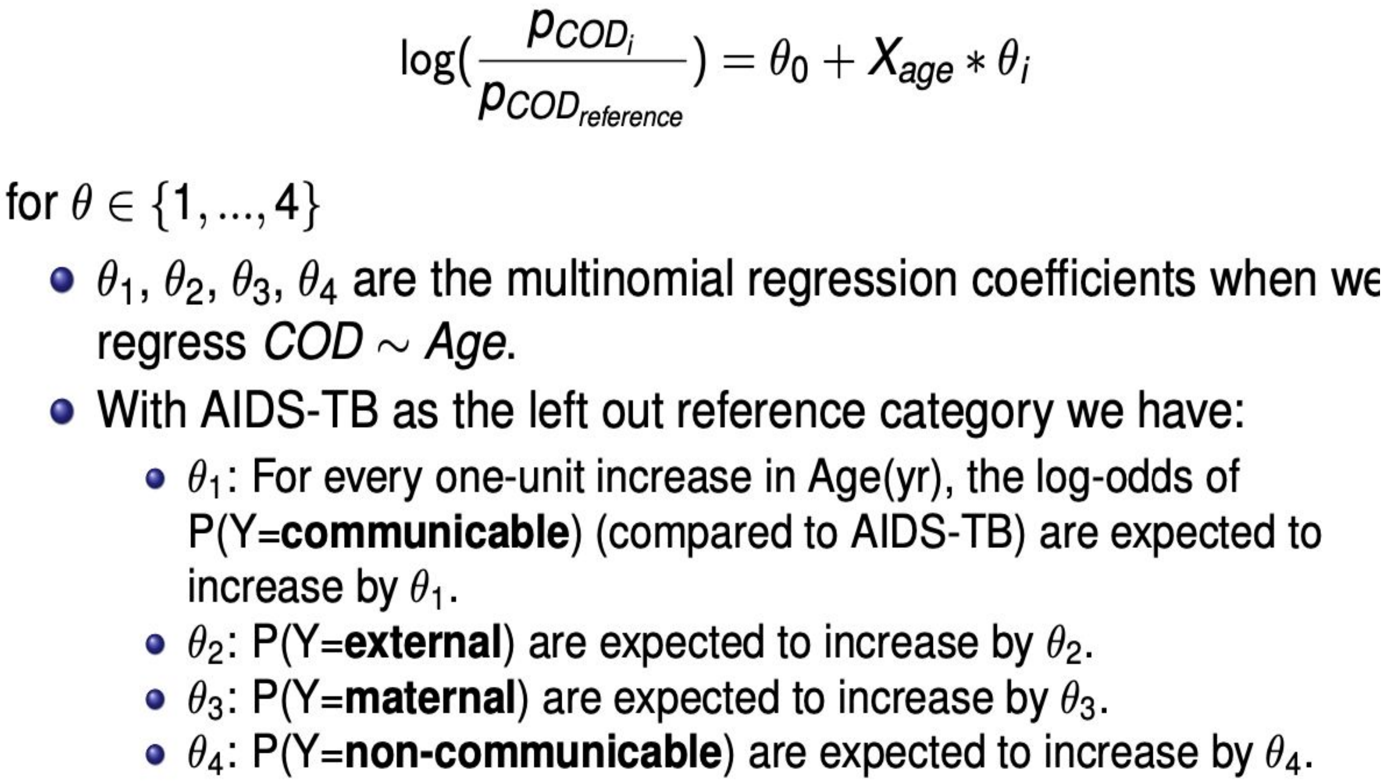


narrative	gs_cod	prediction
respondent thanked for being visited	aids-tb	The narrative does not provide enough information to determine a cause of death.
client had no additional point	non-communicable	The narrative does not provide enough information to determine the appropriate label.
the client thanked for service which provided in the hospital_x000d__x000d_\nthe client transfer death certificate to their original home [place]	non-communicable	The narrative does not provide enough information to determine the cause of death.
the client thanked for the service	communicable	The narrative does not provide information related to any of the labels.
no comment	communicable	The narrative does not provide enough information to determine the cause of death.

- GPT-4 fails to classify 1503 of the 6763 cases. These 1503 text narratives contain no useful information.

How does Age (X) vary with Cause of Death (y)?

multinomial logistic regression:



$$\log(\frac{p_{COD_i}}{p_{COD_{reference}}}) = \theta_0 + X_{age} * \theta_i$$

for $\theta \in \{1, ..., 4\}$

- θ_1, θ_2, θ_3, θ_4 are the multinomial regression coefficients when we regress $COD \sim Age$.
- With AIDS-TB as the left out reference category we have:
    - θ_1: For every one-unit increase in Age(yr), the log-odds of P(Y=**communicable**) (compared to AIDS-TB) are expected to increase by θ_1.
    - θ_2: P(Y=**external**) are expected to increase by θ_2.
    - θ_3: P(Y=**maternal**) are expected to increase by θ_3.
    - θ_4: P(Y=**non-communicable**) are expected to increase by θ_4.