Valid Inference With Predictions From Narratives

Shuxian Fan, Adam Visokay, Kentaro Hoffman, Li Liu, Stephen Salerno, Tyler H, McCormick and Jeff Leek

UNIVERSITY of WASHINGTON
CENTER FOR STUDIES IN DEMOGRAPHY & ECOLOGY
CENTER for STATISTICS and the SOCIAL SCIENCES
eScience Institute
IMPRS for Population, Health and Data Science
INTERNATIONAL MAX PLANCK RESEARCH SCHOOL
JOHNS HOPKINS UNIVERSITY
Fred Hutch Cancer Center

MOTIVATION

Verbal Autopsies (**VA**) are interviews used to predict cause of death (**COD**) in low resource settings.

Statistical Inference using **AI predictions** instead of **ground truth** labels will produce biased results.

Question: How can we perform valid statistical inference even with AI predicted causes of death?

DATA

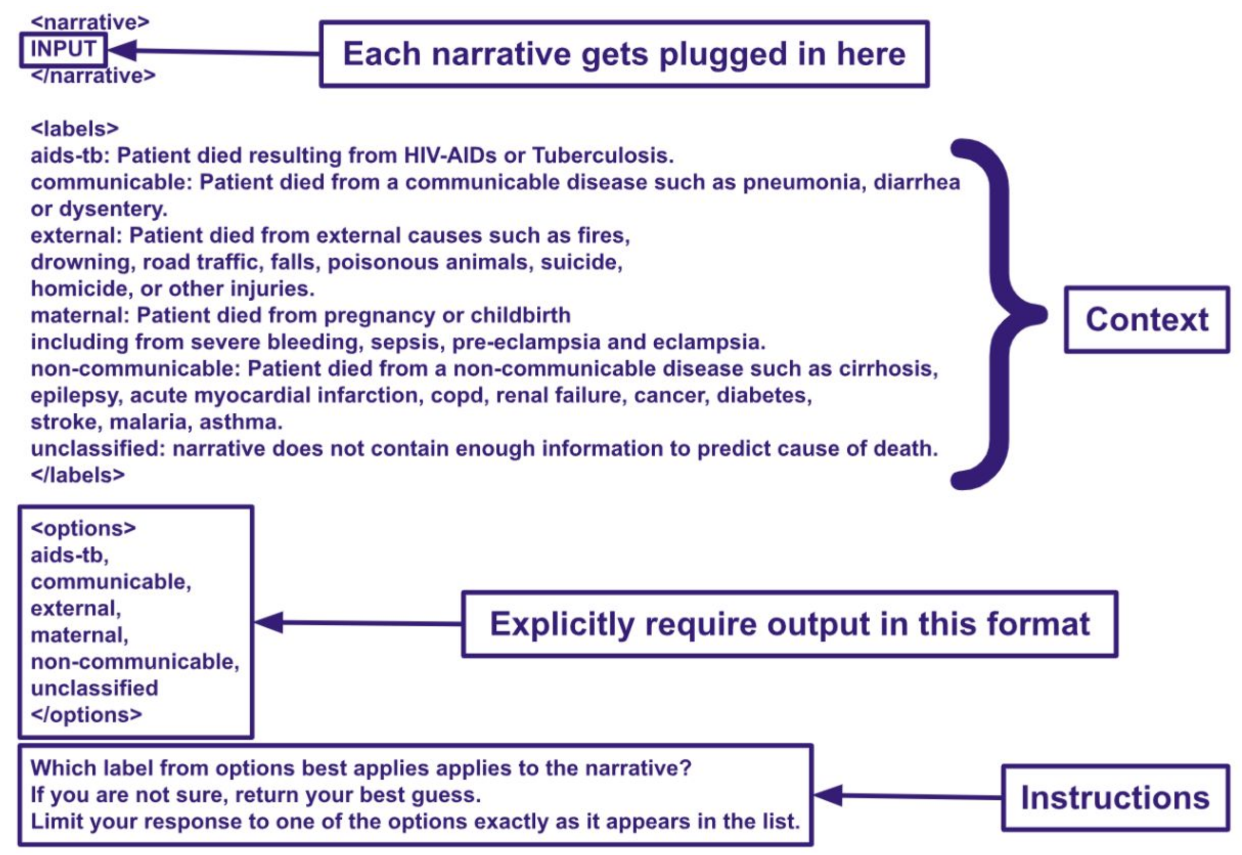
Population Health Metrics Research Consortium

- 2005, in 6 sites, from 4 countries.
- Adults only, 5 cause of death labels.
- n = 6763 total observations.
- This dataset uniquely has verbal and real autopsies which allows us to create calibrated prediction models.
- We use traditional machine learning models as well as GPT-4.

TRADITIONAL NLP

- Models: BERT, SVM, KNN and NB.
- Top **F1-scores** in **[0.58-0.67]**
- Fast and cheap predictions.

PROMPT ENGINEERING



Traditional NLP only outputs training classes, LLM output has no constraint.

LLM outputs may be correct, but hard to validate at scale, or entirely unuseful. For example:

1) "Pain in back 🤔 "
2) "They thanked the interviewer."

Returning "unclassified" genuinely uninformative narratives is something traditional NLP methods cannot do.

GPT PREDICTIONS

- **F-1 score** of **0.45**. with mis-classified labels. 🥇
- Drop "unclassified" predictions, **F1-score** of **0.75**.
- Better than traditional NLP, but **cost $3,000**. **$$$**

SCIENTIFIC MODEL

Inference on Predicted Data, we use multinomial logistic regression to infer association between **Age** and **COD**:

$$log\left(\frac{p_{COD_i}}{p_{COD_{ref}}}\right) = \theta_i Age$$

θ_i is change in log-odds of person i being classified with COD_i relative to the reference cause AIDS-TB

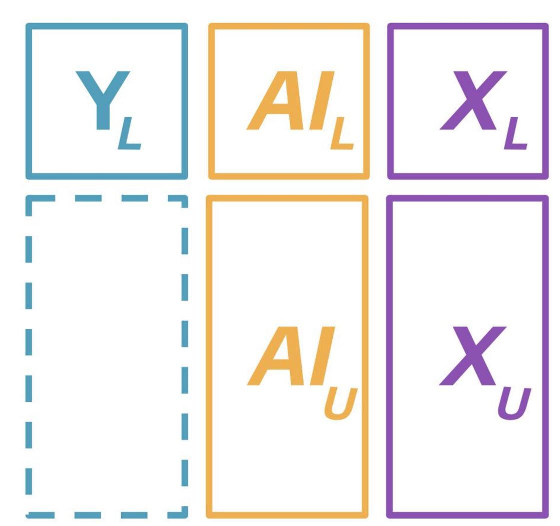
LOSS FUNCTION

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

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

STATISTICAL CORRECTION

We use *some* **labeled** but *mostly* **AI Predicted** data to create a **correction factor** which we use to recover corrected θ_i



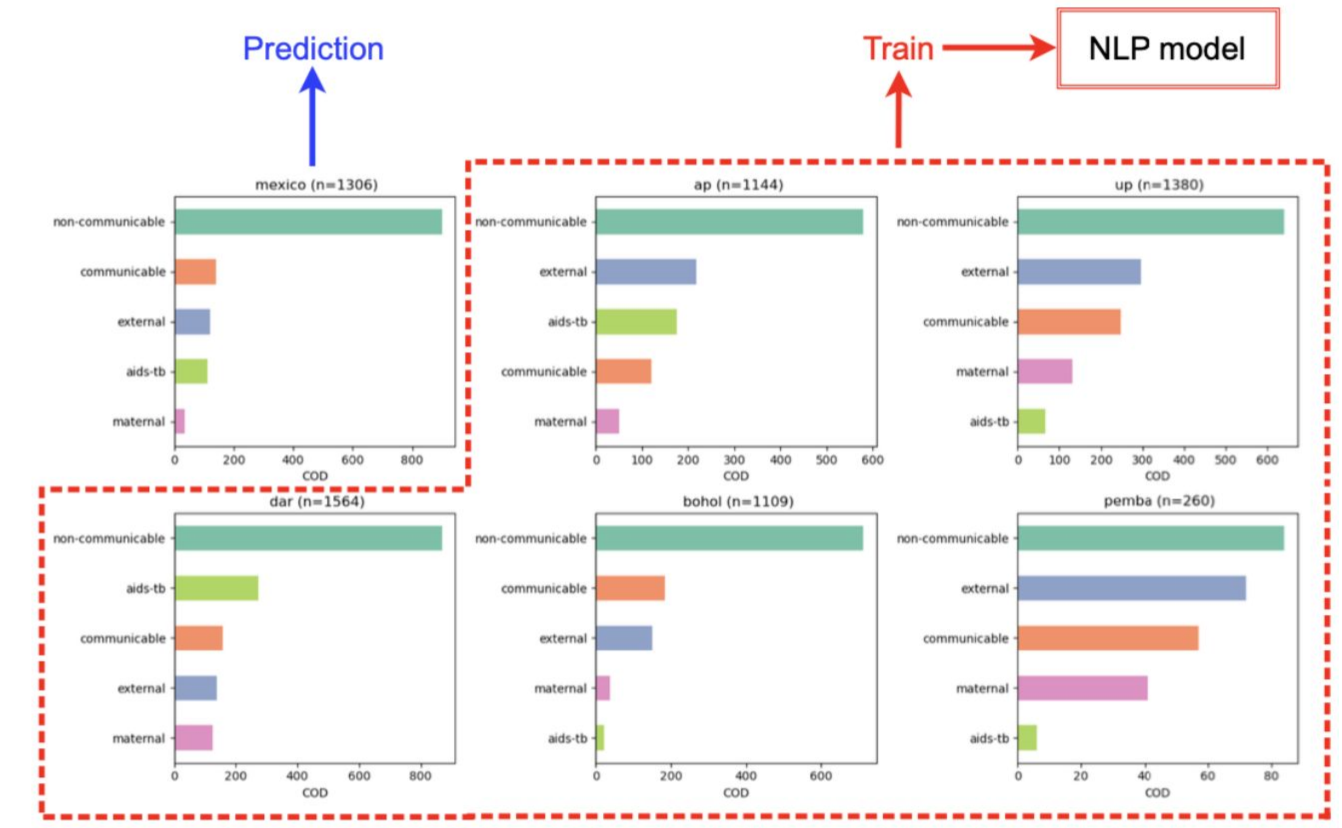
20% of our data have
- **True** Labels (Y_L),
- **AI** Labels (AI_L),
- Observed **Age** (X_L).

80% with AI_U and X_U.

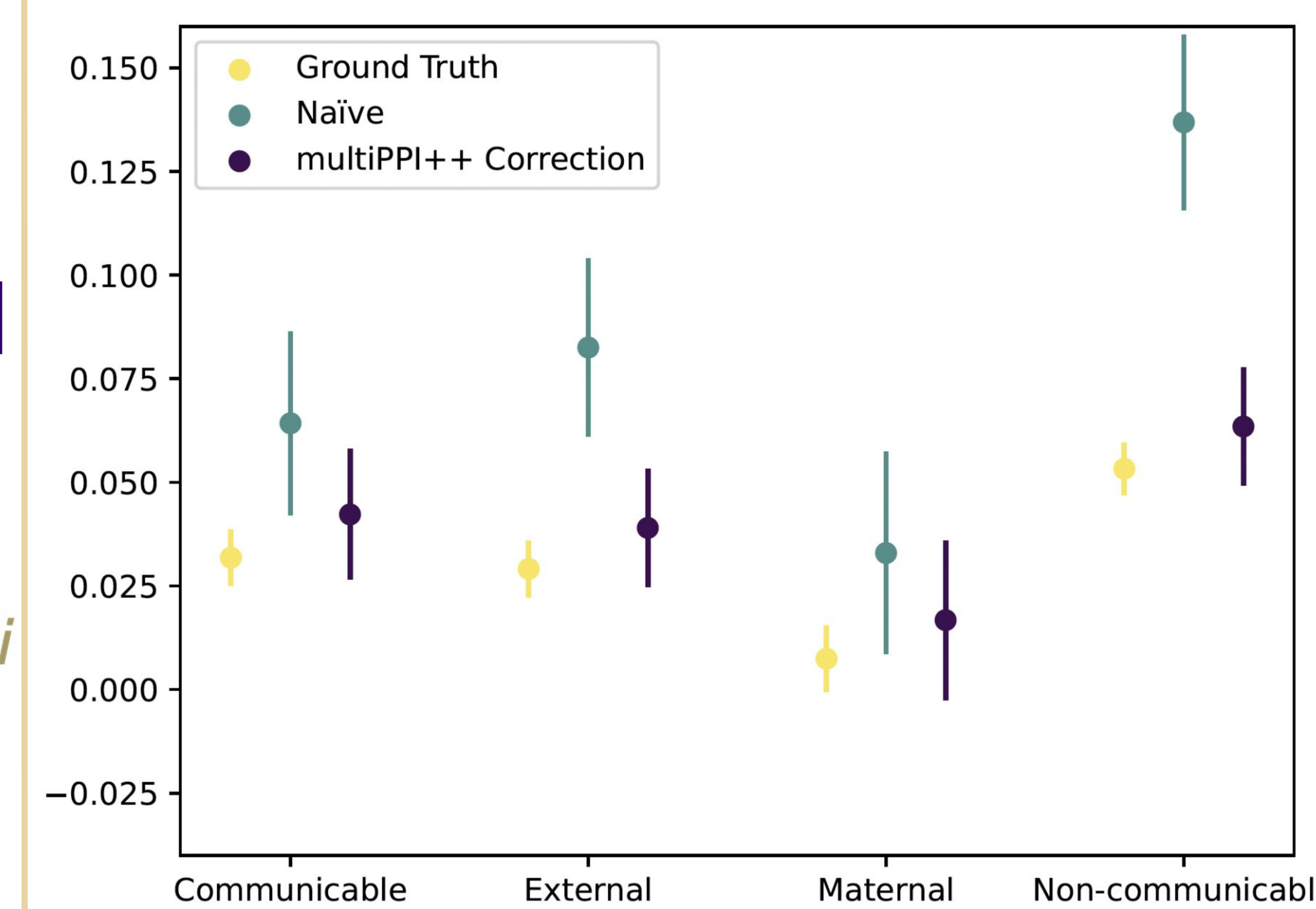
Step 1	$\hat{\theta}^{AI}$: $AI_U \sim X_U$
Step 2	θ^{True} : $Y_L \sim X_L$
Step 3	$\hat{\Delta}$: $(Y_L - AI_L) \sim X_L$
Step 4	$\hat{\theta}^{corrected} = \hat{\theta}^{AI} + \hat{\Delta}$

TRANSPORTABILITY

Leave-one-out train/predict for each site.



CORRECTED ESTIMATES



Use labeled and unlabeled data for valid inference on predicted data!

REFERENCES

- Murray et al PHMRC (2011)
- Egami et al (2023)
- Surek-Clark (2020)
- Wang et al (2020)
- Angelopoulos et al (2023a/b)

COLM

Full Paper