Valid Inference Using Verbal Autopsy Narratives

Joint work with Trinity Fan, Kentaro Hoffman, Li Liu, Stephen Salerno, Tyler McCormick and Jeff Leek

eScience Institute

CENTER for STATISTICS and the SOCIAL SCIENCES

W ADAM VISOKAY

DEPARTMENT OF SOCIOLOGY

MOTIVATION

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

Inference using **AI predictions** instead of **ground truth** labels will produce **biased** point estimates and misleadingly **narrow uncertainty.**

How can we correct for this and perform valid inference?

DATA

Population Health Metrics Research Consortium

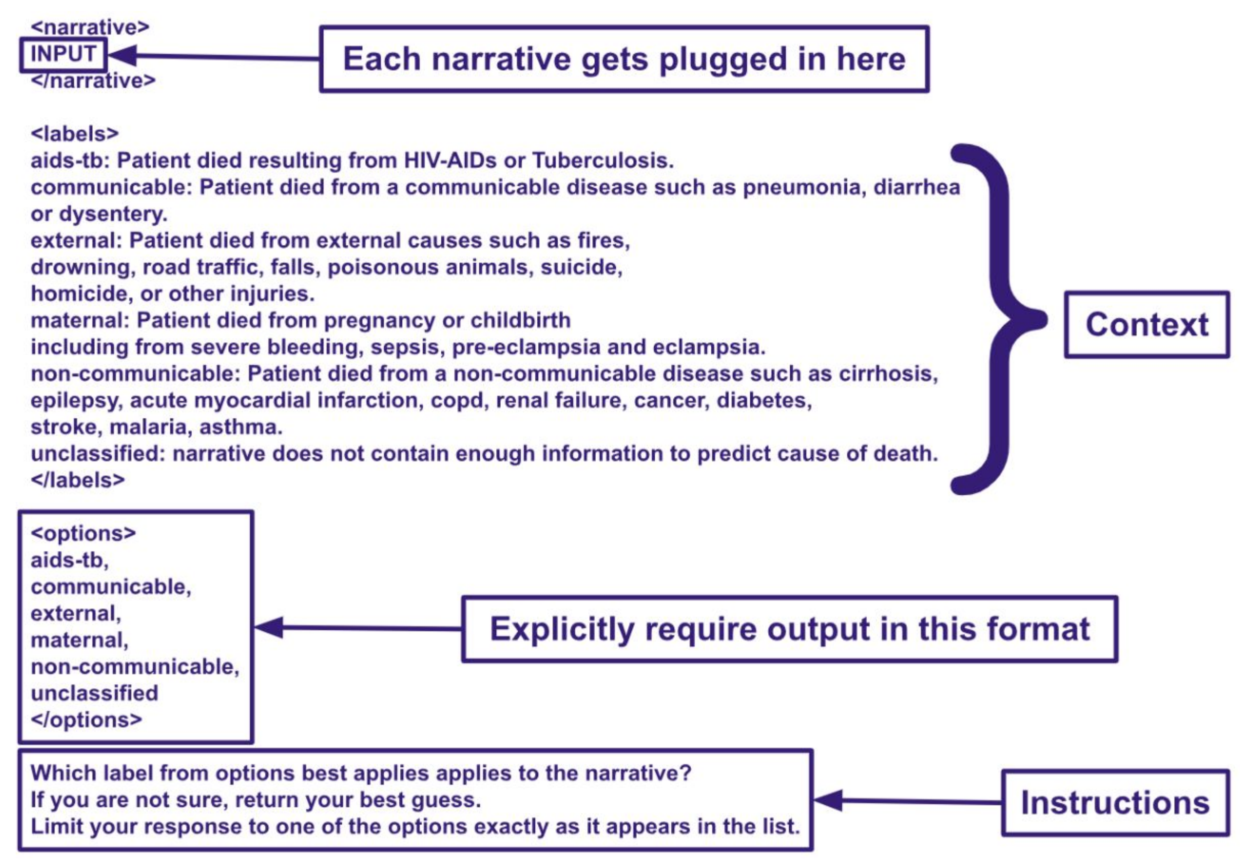
- **2005**, in **6** sites, from **4** countries.
- Adults only, **5** cause of death labels.
- n = **6763** total observations.
- VA **text narratives** with **Gold Standard** labels from traditional autopsies.

We use **NLP** with traditional machine learning and a state of the art **LLM.**

TRADITIONAL NLP

- BERT with Bag of Words representation, SVM, KNN and NB.
- Achieves low **F1-scores** ranging from **0.58** to **0.67**, but **cheaper** to compute.

GPT-4 PROMPT



- Traditional NLP returns only classes from training set.
- **LLM** output is **unconstrained**.

GPT-4: Pain in back 🤔

- **OOPS...**
- But wait! This "limitation" is actually **helpful**!
- GPT-4 says **THIS NARRATIVE IS USELESS!** Good!

GPT PREDICTIONS

- GPT-4 achieves lousy **F-1 score of 0.45** with mis-classified labels.
- **BUT,** when we drop the unclassified predictions, GPT-4 **F1-score is 0.75**.
- GPT-4 outperforms traditional NLP, but it **cost us ~ $3,000!!!**

$$$

INFERENCE TASK

Multinomial logistic regression to infer association between **Age** and **COD**:

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

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

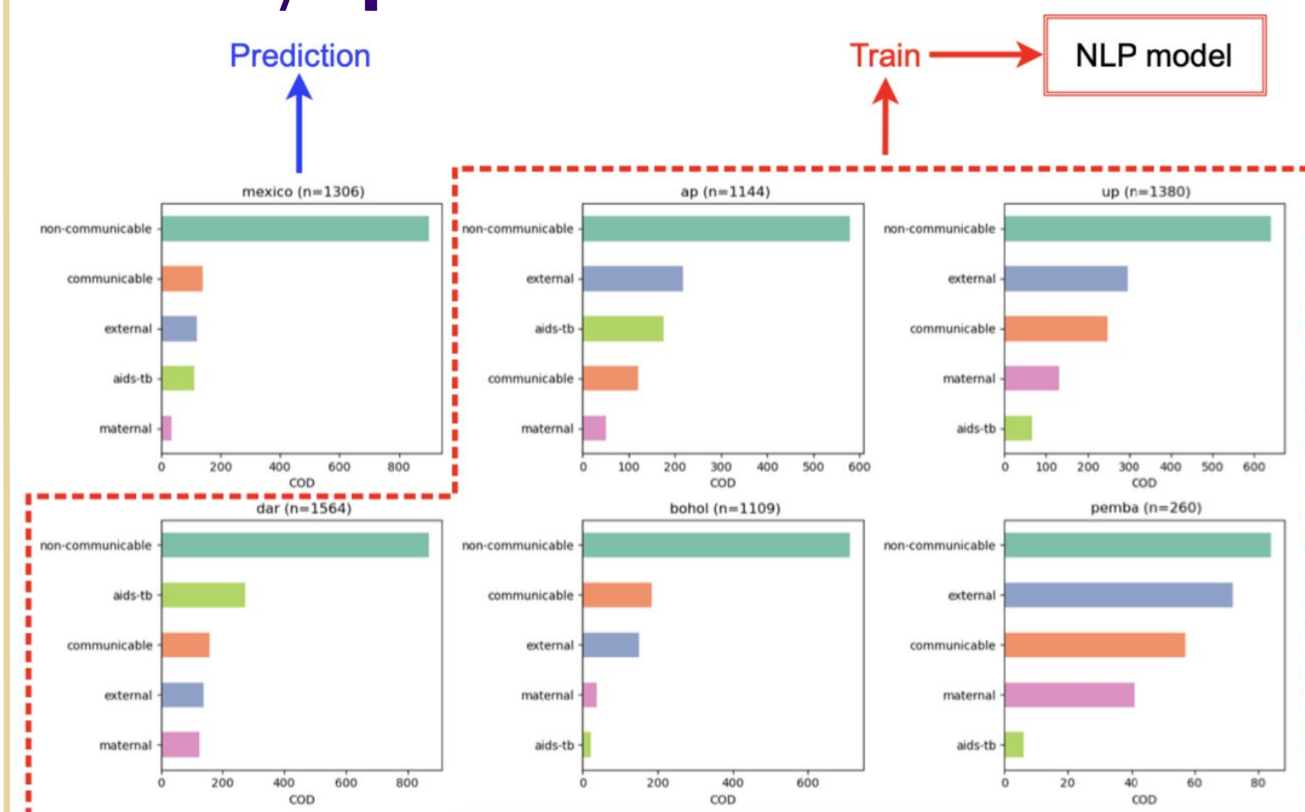
MODELING ERROR

AI predicted labels are a **best guess**.

$$\widehat{COD}_{AI} \neq COD_{True\ Label}$$

TRANSPORTABILITY

Training in **Domain A** doesn't always **predict** well in **Domain B**.



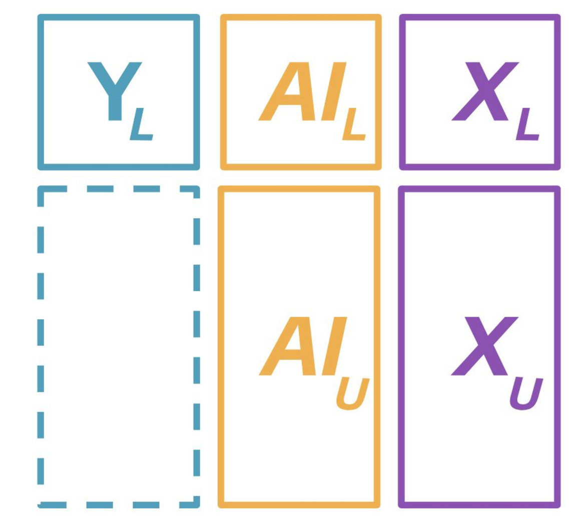
$$\text{ARGMIN}\{\theta \in \mathbb{R}^{d}\}$$

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

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

STATISTICAL CORRECTION

We use *some* **Labeled** with *mostly* **AI Predicted** data to create a **correction factor** which we use to recover true θ_i.



20% of our data have

- **True** Labels (Y_L),
- **AI** Labels (AI_L),
- Observed **Age** (X_L).

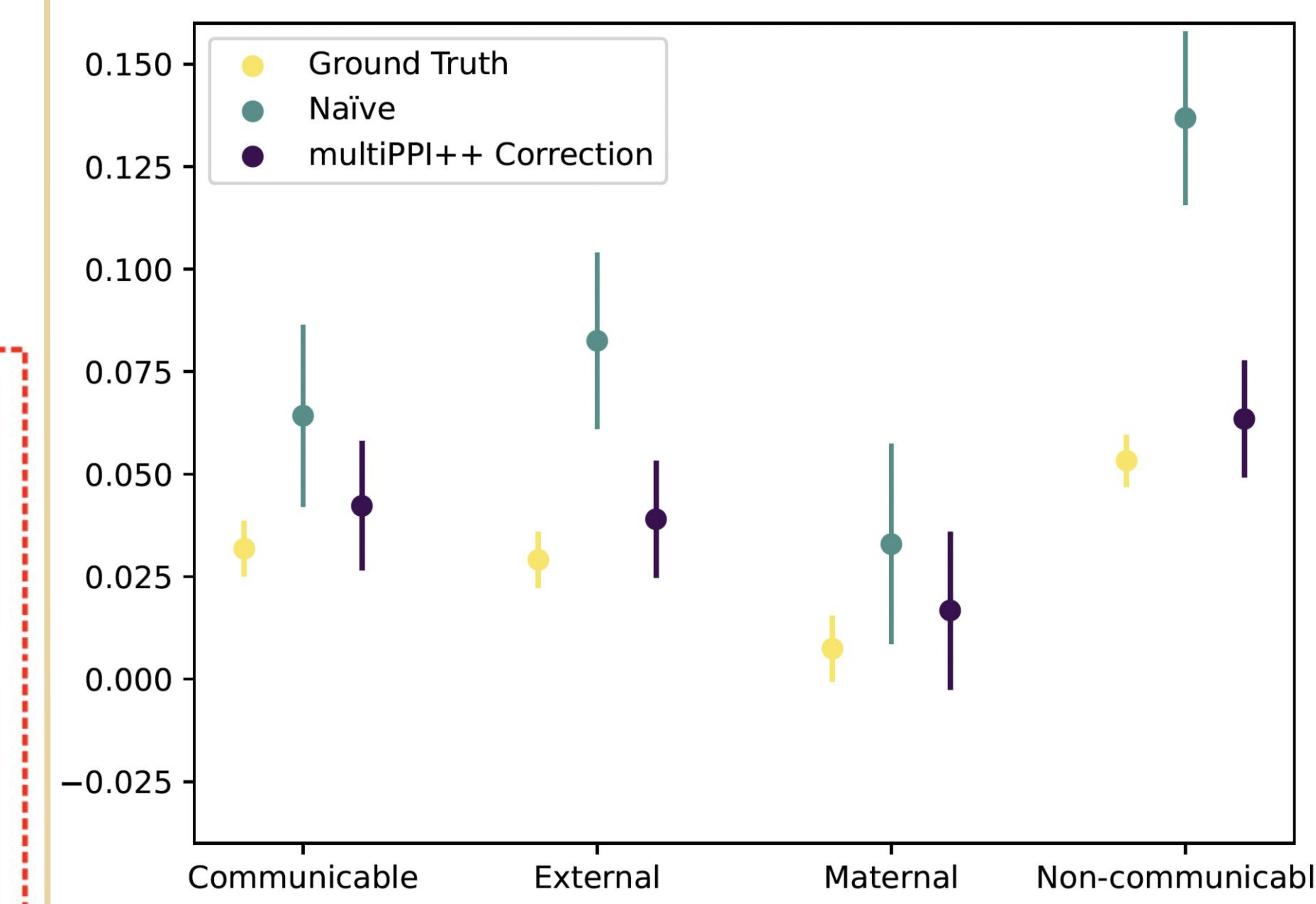
80% with AI_U and X_U.

Step **1** $\quad \hat{\theta}^{AI}$: AI_U ~ X_U

Step **2** $\quad \theta^{True}$: Y_L ~ X_L

Step **3** $\quad \hat{\Delta}$: $(Y_L - AI_L)$ ~ X_L

Step **4** $\quad \hat{\theta}^{corrected} = \hat{\theta}^{AI} + \hat{\Delta}$



REFERENCES

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

Full Paper