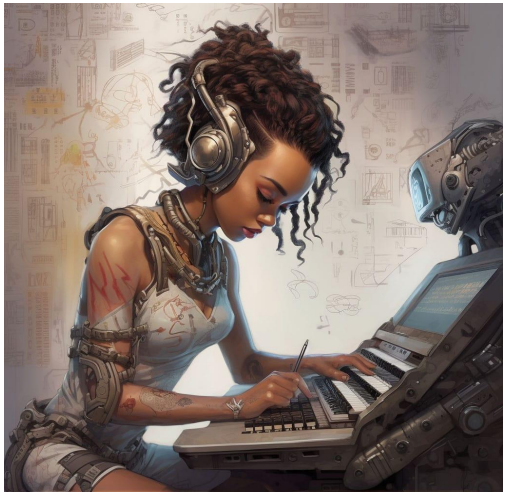
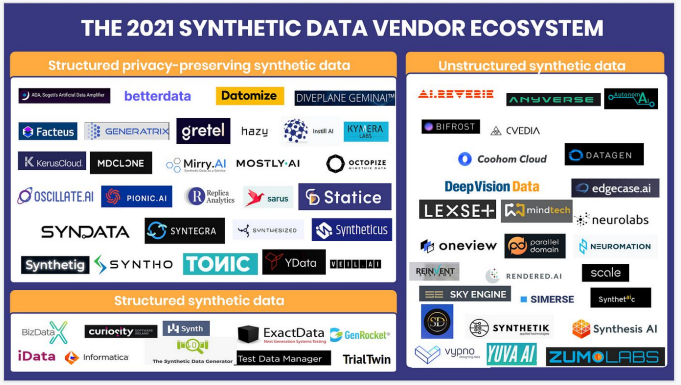
Valid Inference After AI/ML Prediction

Adam Visokay and Kentaro Hoffman
xD at the US Census Bureau
July 10, 2024

Generative AI: A wellspring of Synthetic Data?





The ability to mimic complex human generated structures (ex speech) has shown tremendous potential.

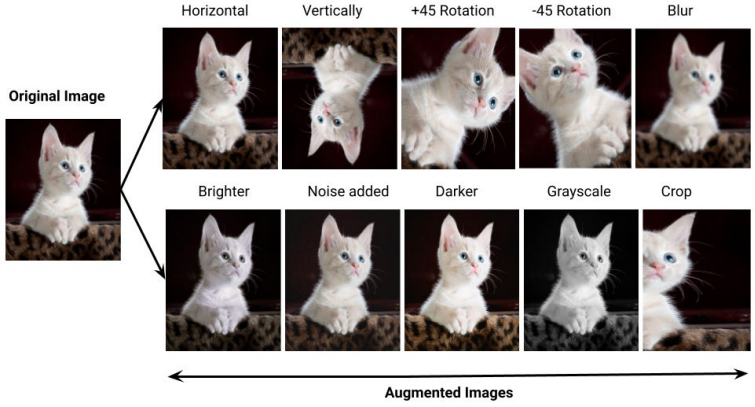
Synthetic Data for Staticians: An Opportunity and a Danger

Opportunity

- Reduce data collection costs
- Improve robustness

Danger

- Propagate existing bias
- “Black box” nature limits interpretability
- Potential for model collapse



arXiv > cs > arXiv:2402.07712

Computer Science > Machine Learning

[Submitted on 12 Feb 2024 (v1), last revised 30 Apr 2024 (this version, v2)]

Model Collapse Demystified: The Case of Regression

Elvis Dohmatob, Yunzhen Feng, Julia Kempe

Our Perspective

Generative AI is exceedingly common in research pipelines throughout the social sciences, public health, demography, and beyond.

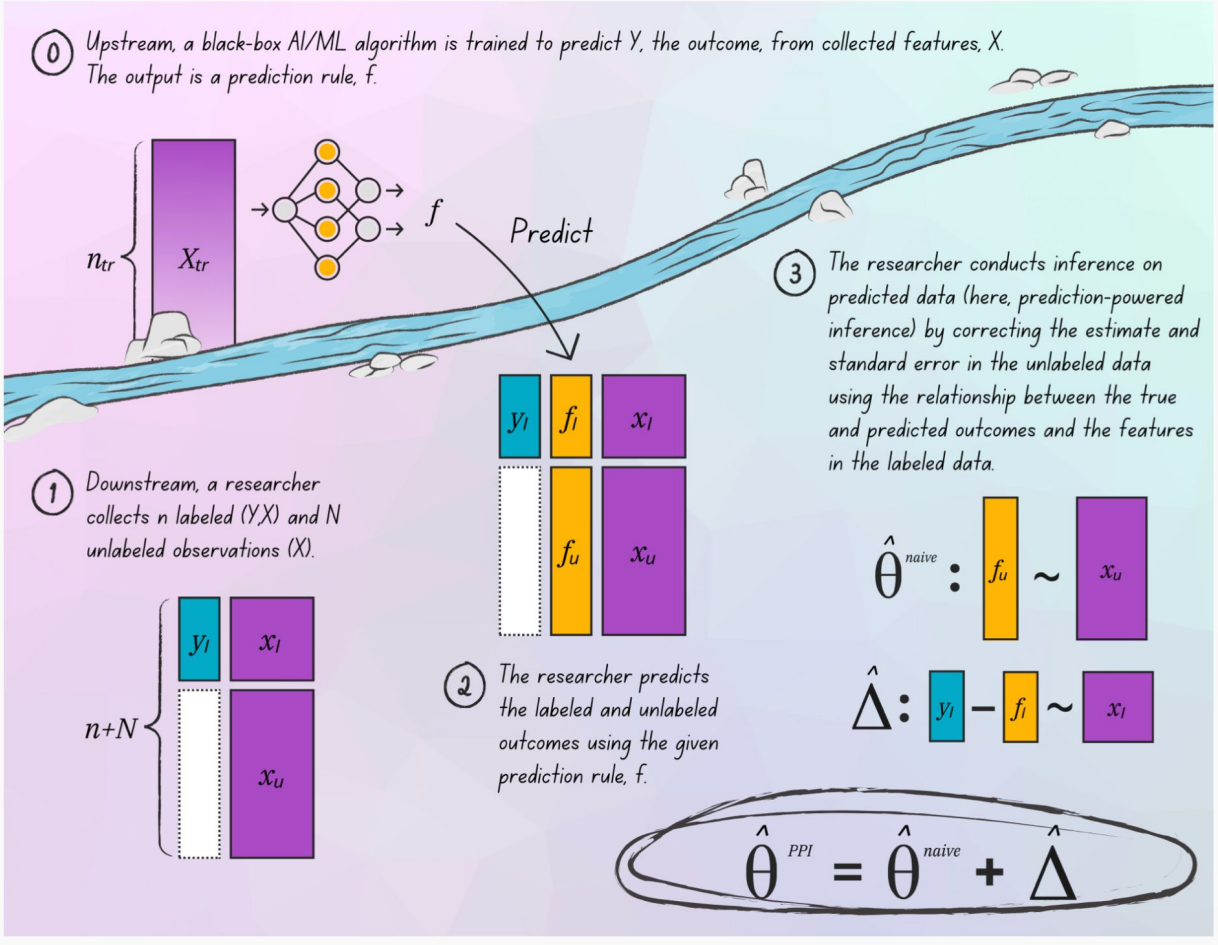
We should create statistical methods which work with AI generated data.

The Setup:

- Observing a specific outcome of interest, **Y**, is expensive.
- A scientist augments an existing dataset [**X** and **Y**] with AI generated synthetic outcomes [**X** and **f(X)**] and wants to use both **Y** and **f(X)** to learn about the statistical relationship between **X** and **Y**.

We call this **Inference on Predicted Data (IPD)**

Overview of Inference on Predicted Data



PostPI: Siruo Wang, Tyler H. McCormick, & Jeffrey T. Leek (**2020**). Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences, 117(48), 30266-30275.*

PPI: Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, & Tijana Zrnic. (**2023**). Prediction-Powered Inference.

PPI++: Anastasios N. Angelopoulos, John C. Duchi, & Tijana Zrnic. (**2024**). PPI++: Efficient Prediction-Powered Inference.

Naoki Egami, Musashi Jacobs-Harukawa, Brandon M Stewart, and Hanying Wei. (**2023**) Using large language model annotations for valid downstream statistical inference in social science: Design-based semi-supervised learning.

Jiacheng Miao, Xinran Miao, Yixuan Wu, Jiwei Zhao, and Qiongshi Lu. (**2023**) Assumption-lean and data-adaptive post-prediction inference.

PPI: Mean Estimation

n- sample size of labeled data (X,Y)
N- Sample size of AI-generated data (X,f(X))

Algorithm 1 Prediction-powered mean estimation

Input: labeled data (X, Y), unlabeled features $\widetilde{X}$, predictor f, error level $\alpha \in (0, 1)$

1: $\hat{\theta}^{\text{PP}} \leftarrow \tilde{\theta}^f - \hat{\mathbf{\Delta}} := \frac{1}{N}\sum_{i=1}^{N} f(\widetilde{X}_i) - \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)$ ▷ prediction-powered estimator

2: $\hat{\sigma}_f^2 \leftarrow \frac{1}{N}\sum_{i=1}^{N}(f(\widetilde{X}_i) - \tilde{\theta}^f)^2$ ▷ empirical variance of imputed estimate

3: $\hat{\sigma}_{f-Y}^2 \leftarrow \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i - \hat{\mathbf{\Delta}})^2$ ▷ empirical variance of empirical rectifier

4: $w_\alpha \leftarrow z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_{f-Y}^2}{n} + \frac{\hat{\sigma}_f^2}{N}}$ ▷ normal approximation

Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\text{PP}} = \left(\hat{\theta}^{\text{PP}} \pm w_\alpha\right)$

- Produces unbiased estimates of theta
- Reduces size of the confidence interval compared to using just real data
- Robust to choice of f
- Requires no retraining/fine tuning of AI model [Off the shelf]

Image from: Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, & Tijana Zrnic. (**2023**). Prediction-Powered Inference.

PPI: Mean Estimation

n- sample size of labeled data (X,Y)
N- Sample size of AI-generated data (X,f(X))

Algorithm 4 Prediction-powered linear regression

Input: labeled data (X, Y), unlabeled features $\widetilde{X}$, predictor f, coefficient $j^* \in [d]$, error level $\alpha \in (0, 1)$

$\hat{\theta}^{\text{PP}} \leftarrow \tilde{\theta}^f - \hat{\boldsymbol{\Delta}} := \widetilde{X}^\dagger f(\widetilde{X}) - X^\dagger(f(X) - Y)$ ▷ prediction-powered estimator

$\tilde{\Sigma} \leftarrow \frac{1}{N}\widetilde{X}^\top \widetilde{X}$, $\tilde{M} \leftarrow \frac{1}{N}\sum_{i=1}^{N}(f(\widetilde{X}_i) - \widetilde{X}_i^\top \tilde{\theta}^f)^2 \widetilde{X}_i \widetilde{X}_i^\top$

$\tilde{V} \leftarrow \tilde{\Sigma}^{-1}\tilde{M}\tilde{\Sigma}^{-1}$ ▷ "sandwich" variance estimator for imputed estimate

$\Sigma \leftarrow \frac{1}{n}X^\top X$, $M \leftarrow \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i - X_i^\top \hat{\boldsymbol{\Delta}})^2 X_i X_i^\top$

$V \leftarrow \Sigma^{-1} M \Sigma^{-1}$ ▷ "sandwich" variance estimator for empirical rectifier

$w_\alpha \leftarrow z_{1-\alpha/2}\sqrt{\frac{V_{j^*j^*}}{n} + \frac{\tilde{V}_{j^*j^*}}{N}}$ ▷ normal approximation

Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\text{PP}} = \left(\hat{\theta}_{j^*}^{\text{PP}} \pm w_\alpha\right)$

- Produces unbiased estimates of theta
- Reduces size of the confidence interval compared to using just real data
- Robust to choice of f
- Requires no retraining/fine tuning of AI model [Off the shelf]

Image from: Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, & Tijana Zrnic. (**2023**). Prediction-Powered Inference.

PPI: Objective Function

$$L(\theta) := \mathbb{E}[\ell_\theta(X, Y)] \quad \text{and} \quad L^f(\theta) := \mathbb{E}[\ell_\theta(X, f(X))].$$

The starting point of PPI and our approach is the recognition that $\mathbb{E}[\ell_\theta(X, f(X))] = \mathbb{E}[\ell_\theta(\widetilde{X}, f(\widetilde{X}))] = L^f(\theta)$, so that the "rectified" loss,

$$L^{\mathrm{PP}}(\theta) := L_n(\theta) + \widetilde{L}^f_N(\theta) - L^f_n(\theta),$$

where

$$L_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i, Y_i), \quad L^f_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i, f(X_i)) \quad \text{and} \quad \widetilde{L}^f_N(\theta) := \frac{1}{N}\sum_{i=1}^{N} \ell_\theta(\widetilde{X}_i, f(\widetilde{X}_i)),$$

Efficient algorithms for estimation exist when L is convex

Image from: Anastasios N. Angelopoulos, John C. Duchi, & Tijana Zrnic. (**2024**). PPI++: Efficient Prediction-Powered Inference.

Real world example: Verbal Autopsies

Fewer than ⅓ of deaths are assigned a cause at time of death (WHO Report, 2017).

Structured Interviews with family or friends of the deceased known as Verbal Autopsies (VA) serve as a cheap proxy to determine COD in resource-scarce settings.

Setup

Imagine you are a public health researcher in a developing country. You are given a budget to study the relationship between COD **[Y]** and age **[X].**

Recently, a research group has trained an AI model on other developing countries which can predict a COD **[f(X)]** given a Verbal Autopsy interview.

“the deceased had been burnt and died within 1.5 hours of the accident”

COD: **Fires**

ICD10: **External**

The AI model isn’t 100% accurate and it was trained on a different country.

But surely it’s gotta be helpful for something….

Setup

As the researcher, you can either:

1. Fly out a coroner to various low resource areas to conduct traditional autopsies. **Real data collection: (X,Y)**
   a. Most expensive, but unbiased inference.
2. Conduct Verbal Autopsies and then use the NLP-AI model to predict the COD from the interviews. **AI-generated data: (X,f(X))**
   a. Cheapest, but the inference is biased and .
3. Combination of **both, with a PPI correction**: Conduct some traditional autopsies (20%), but mostly use VA + NLP/AI predictions (80%).
   a. Cheap, with debiased inference.

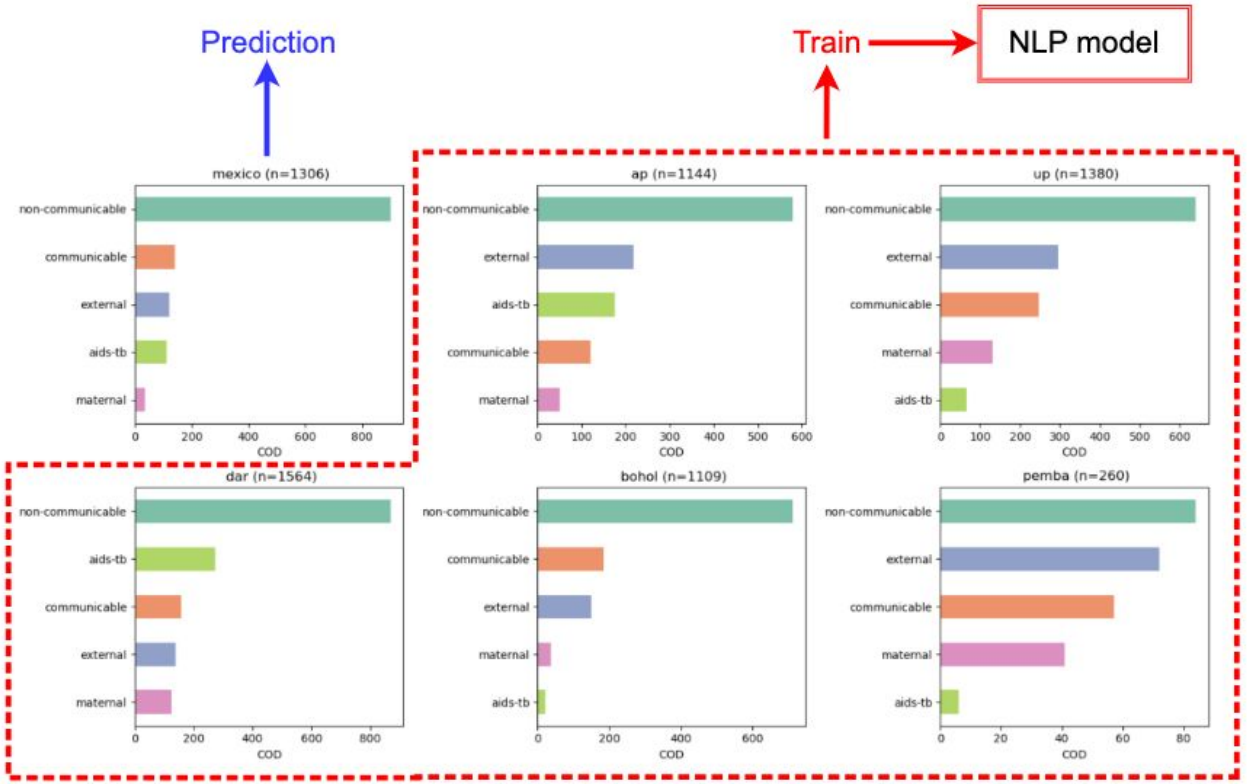
Experimental Design

Imagine the researcher can afford

> 20% [261 samples] traditional autopsies and

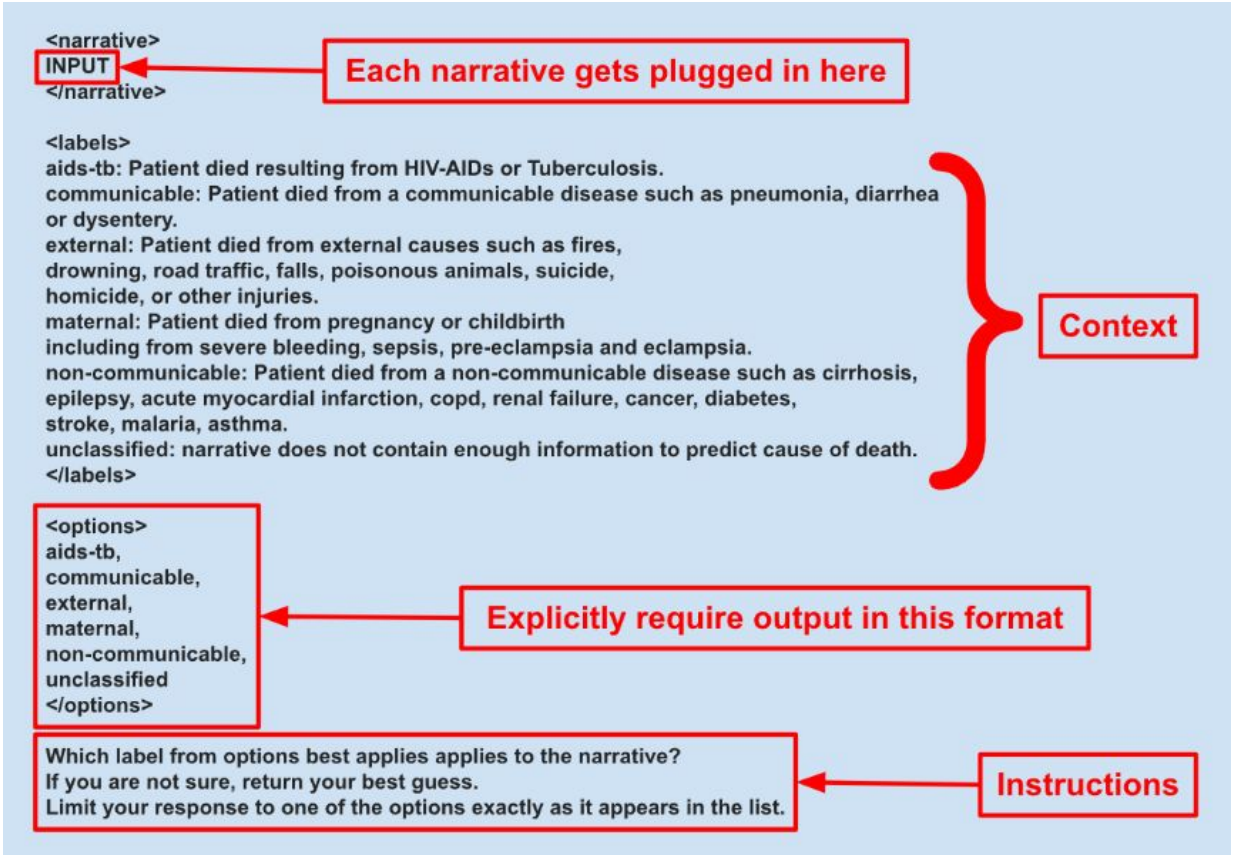
> 80% [1044 samples] verbal autopsies

Transportability Challenge

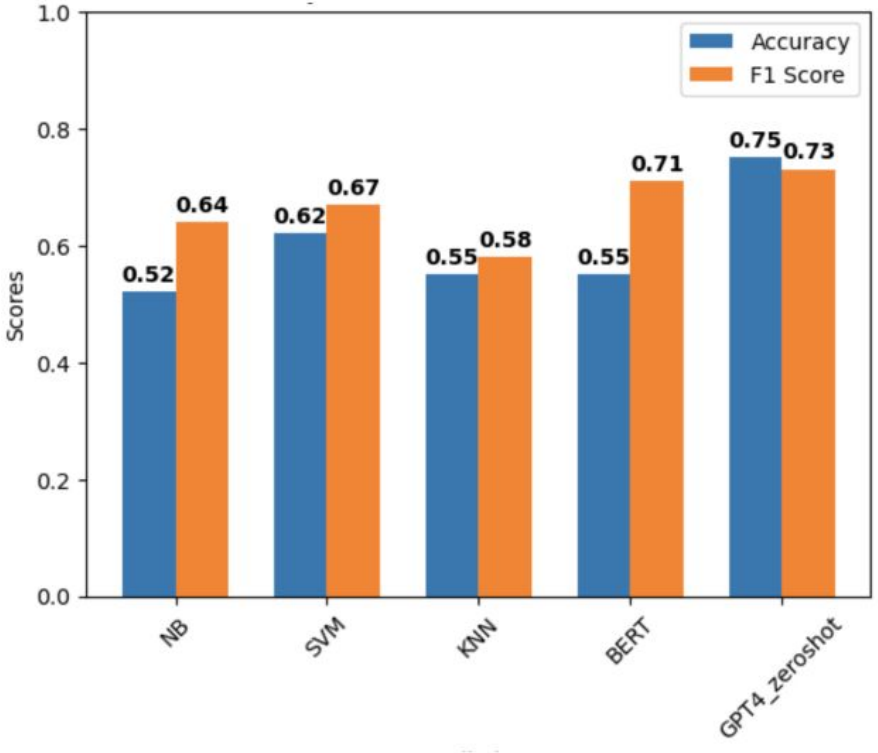


- The other 5 locations were used to train an NLP model
- COD distributions vary wildly due to epidemiological conditions

GPT-4 Zero Shot Prompt



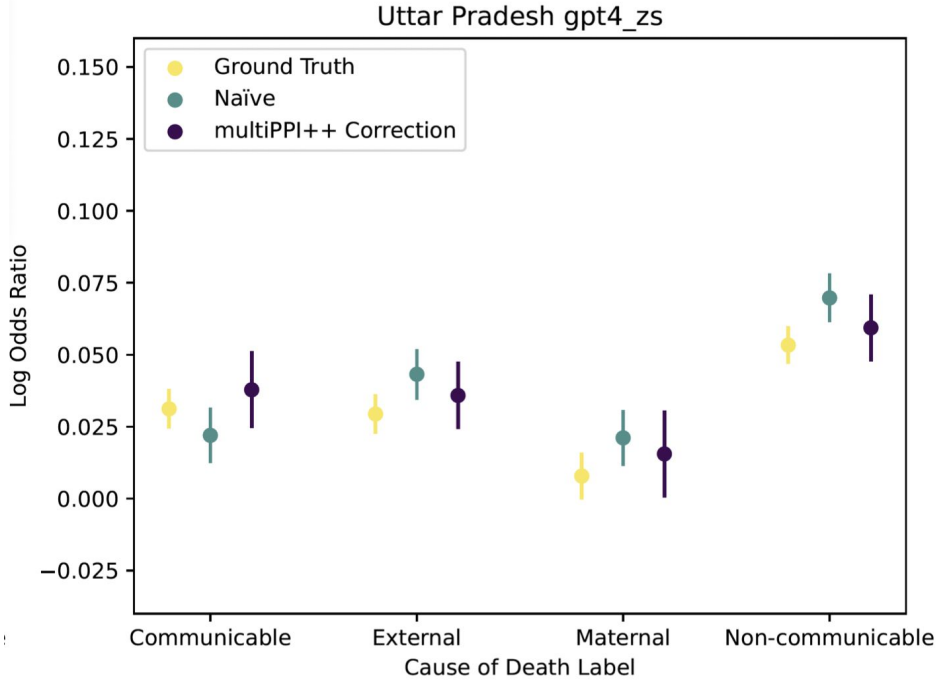
NLP Predictions



We experimented with several NLP prediction models including bag of words + NB, SVM, KNN and deep learning models BERT and GPT-4.

GPT-4 yields the best results with an Accuracy of 0.75 and an F1-Score of 0.73.

PPI Corrections for COD ~ Age



Parameter of interest: Log odds ratio between Age and COD. [Aids-tb is baseline]

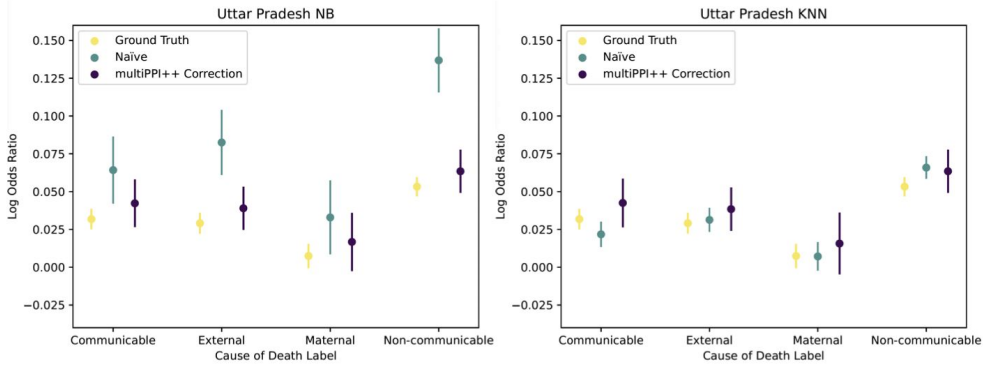
Ground Truth: you magically had the budget to do real autopsies on everyone [Y ~ X]

Naive: if you regress AI generated COD on Age [f(X) ~ Age].

multiPPI++ correction: estimation of log odds using an efficient multinomial version of PPI [[f(X) ~ Age] + rectifier]

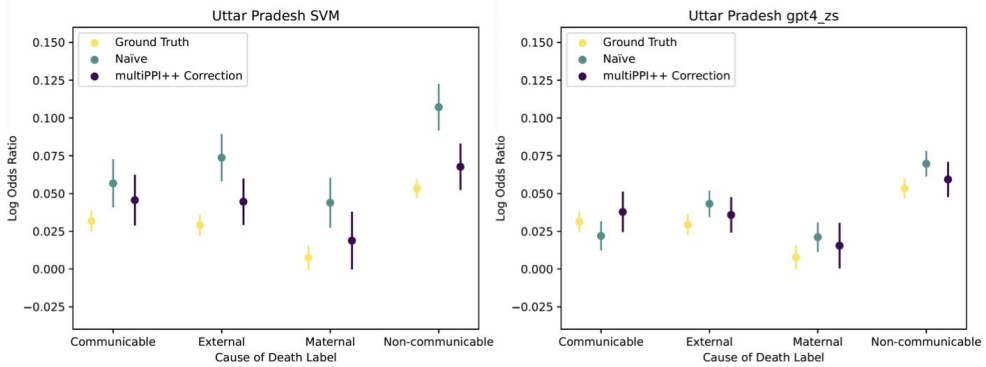
1. PPI correction debiases the point estimate
2. Inflates the uncertainty from using predicted labels

PPI Corrections



(a) BoW with Naïve Bayes

(b) BoW with K-Nearest-Neighbors



(c) BoW with Support Vector Machine

(d) GPT-4-32k

Models with lower accuracy, like Naive Bayes, have more room for PPI correction.

Implications

We can recover accurate parameter estimates combining some ground truth labels with AI-generated proxy labels and a PPI correction.

This opens up several research questions:

- Can we replace in-person interviews with phone interviews? [Our next NIH grant application]
- What implications does this have for experimental design? [How do we do a PPI assisted power calculation? Pragmatic clinical trials? Data collection?]
- What if the AI is trained on survey data with known weights? [NHANES] How do we incorporate these weights to make it extrapolate better?
- Implications for privacy and disclosure? [looking for those who know more]

Our team

Adam Visokay

Trinity Fan

Kentaro Hoffman

Stephen Salerno

Li Liu

Jeff Leek

Tyler McCormick

Thank You!

Full Paper

Github Repo

Appendix

PPI++: Objective Function

$$\hat{\theta}_{\lambda}^{\mathrm{PP}} = \arg\min_{\theta} L_{\lambda}^{\mathrm{PP}}(\theta), \text{ where } L_{\lambda}^{\mathrm{PP}}(\theta) := L_n(\theta) + \lambda \cdot (\widetilde{L}_N^f(\theta) - L_n^f(\theta)).$$

- Lambda \in [0,1] is chosen in a data-adaptive fashion to represent how reliable the AI model is
- Lambda=0 Results in classical Inference [aka ignore AI generated data]
- Closely related to AIPW

$$\begin{aligned}
\hat{\theta}^{\mathrm{AIPW}} &:= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mathbb{E}}[Y \mid X = X_i]) + \frac{1}{N+n}\left(\sum_{i=1}^{n}\hat{\mathbb{E}}[Y \mid X = X_i] + \sum_{i=1}^{N}\hat{\mathbb{E}}[Y \mid X = \widetilde{X}_i]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i)) + \frac{1}{N+n}\left(\sum_{i=1}^{n} f(X_i) + \sum_{i=1}^{N} f(\widetilde{X}_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i + \frac{1}{1+\frac{n}{N}}\left(-\frac{1}{n}\sum_{i=1}^{n} f(X_i) + \frac{1}{N}\sum_{i=1}^{N} f(\widetilde{X}_i)\right),
\end{aligned}$$

Image from: Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, & Tijana Zrnic. (**2023**). Prediction-Powered Inference.

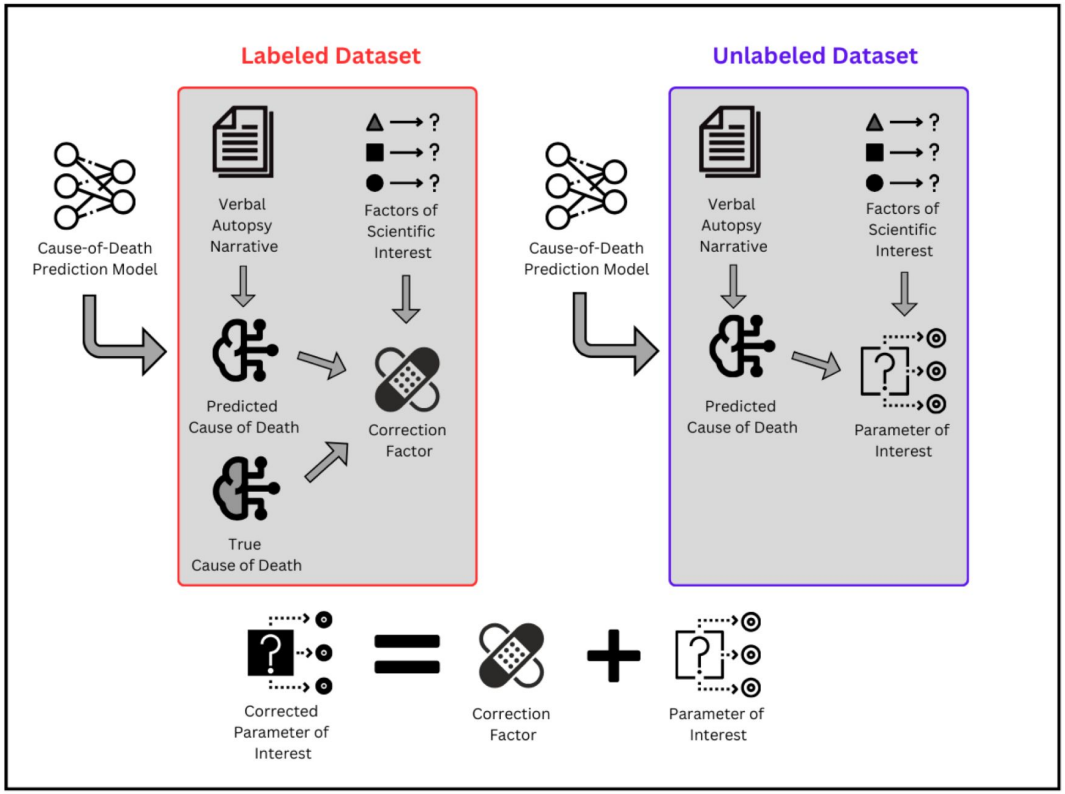
ICD Codes

ICD-10 COD Classification

Mapping 34 PHMRC All-Cause Mortality Labels to Five Broad COD Labels	
All-Cause Mortality Label	**Broad COD Label**
cirrhosis	non-communicable
epilepsy	non-communicable
pneumonia	communicable
copd	non-communicable
acute myocardial infarction	non-communicable
fires	external
renal failure	non-communicable
lung cancer	non-communicable
maternal	maternal
drowning	external
other cardiovascular diseases	non-communicable
aids	aids-tb
other non-communicable diseases	non-communicable
falls	external
road traffic	external
diabetes	non-communicable
other infectious diseases	communicable
tb	aids-tb
suicide	external
other injuries	external
cervical cancer	non-communicable
stroke	non-communicable
malaria	non-communicable
asthma	non-communicable
colorectal cancer	non-communicable
homicide	external
diarrhea/dysentery	communicable
breast cancer	non-communicable
leukemia/lymphomas	non-communicable
poisonings	external
prostate cancer	non-communicable
esophageal cancer	non-communicable
stomach cancer	non-communicable
bite of venomous animal	external

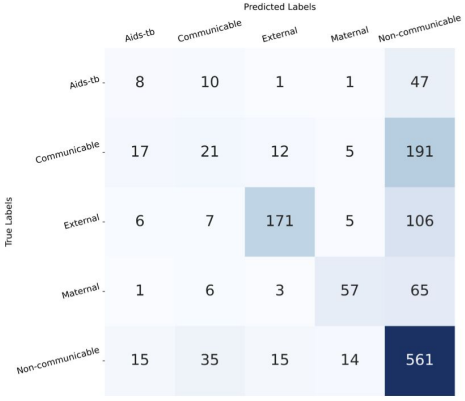
Workflow

The Workflow for Valid Inference Using `multiPPI++` for VA Narratives.

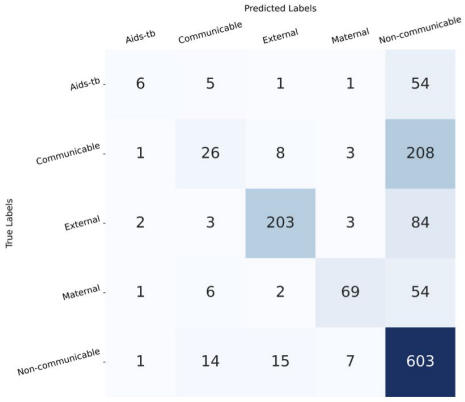


Classification Accuracies

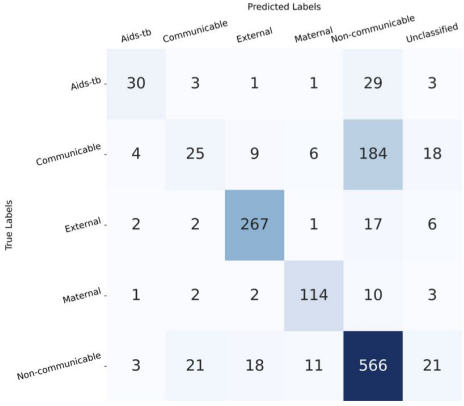
(a) BoW with Naïve Bayes



(b) BoW with K Nearest Neighbors



(c) BoW with Support Vector Machine



(d) GPT-4

MultiPPI

Algorithm 1: `multiPPI++`: Prediction-powered inference for Multinomial Classification

Input: labeled K-category COD data $\{(X_{li}, Y_{li})\}_{i=1}^{n}$, unlabeled data $\{X_{ui}\}_{i=1}^{N}$, NLP model f, significance level $\alpha \in (0,1)$, coefficient index $j \in [d(K-1)]$

1. Optimally select tuning parameter $\hat{\lambda}$ `// set tuning parameter`

2. $\hat{\theta}^{\mathrm{m}}_{\hat{\lambda}} = \underset{\theta \in \mathbb{R}^{d(K-1)}}{\arg\min}\ L^{\mathrm{m}}_{\hat{\lambda}}(\theta)$ `// multiPPI++ estimator`

3. $\hat{H} = \frac{1}{N+n}\left(\sum_{i=1}^{n} \psi''(X_{li}^T \hat{\theta}^{\mathrm{m}}_{\hat{\lambda}}) X_{li} X_{li}^T + \sum_{i=1}^{N} \psi''(X_{ui}^T \hat{\theta}^{\mathrm{m}}_{\hat{\lambda}}) X_{ui} X_{ui}^T\right)$, where
   $\psi(\theta, x) = \log\left(\sum_{k=1}^{K-1} e^{x^T \theta_k}\right), \theta_k \in \mathbb{R}^d$ `// empirical Hessian`

4. $\hat{\Sigma} = \hat{H}^{-1}\left(\frac{n}{N}\hat{V}_f + \hat{V}_\Delta\right)\hat{H}^{-1}$, where

   $\hat{V}_f = \hat{\lambda}^2 \widehat{\mathrm{Cov}}_{N+n}((\psi'(X_{li}^T \hat{\theta}^{\mathrm{m}}_{\hat{\lambda}}) - \hat{Y}_{li}^f))X_{li})$ and
   $\hat{V}_\Delta = \widehat{\mathrm{Cov}}_n((1-\hat{\lambda})(\psi'(X_{li}^T \hat{\theta}^{\mathrm{m}}_{\hat{\lambda}}) + (\hat{\lambda}\hat{Y}_{li}^f - Y_{li})X_{li})$ `// covariance estimator`

Output:
Prediction-powered point estimates $\hat{\theta}^{\mathrm{m}}_{\hat{\lambda}}$ and confidence interval

$\mathcal{C}^{\mathrm{m}}_{\alpha} = \left(\hat{\theta}^{\mathrm{m}}_{\hat{\lambda},j} \pm z_{1-\alpha/2}\sqrt{\hat{\Sigma}_{jj}/n}\right)$ for coordinate j

How

Many modern statistical problems involve using the data you can easily access to ***predict*** the variables you want but cannot measure.

These *predictions* then get used for downstream analysis or data-driven policy decision-making.

Predicted labels are imperfect, and we should account for this.

PHMRC Data

Collection of traditional autopsies and VA for each observation.

n=6763 adults.

Affords us “ground truth” labels which we can use for prediction validation.

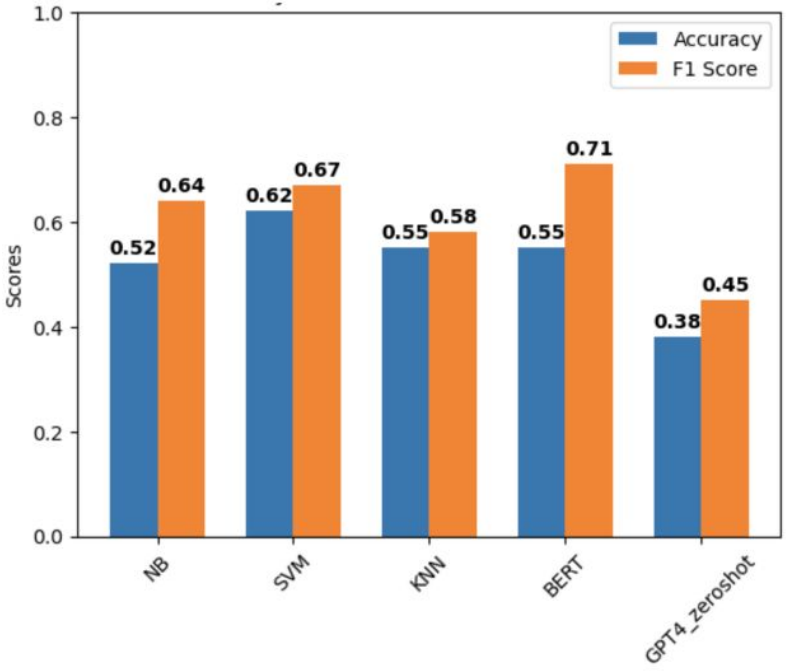
Six sites in four countries.

Collected in 2005.

Experimental Design

- We compare the PPI++ corrected results with the baseline model outputs using the complete dataset to demonstrate the validity of our approach.

NLP Predictions



NLP Predictions - before and after dropping ‘unclassified’