Inference with Predicted Data

Stephen Salerno[1] Jiacheng Miao[2] Awan Afiaz[1,3] Kentaro Hoffman[3] Jesse Gronsbell[4] Jianhui Gao[4] David Cheng[5,6] Anna Neufeld[7] Qiongshi Lu[2] Tyler H. McCormick[3] Jeffrey T. Leek[1,3]

[1] Fred Hutch Cancer Center [2] Univ. of Wisconsin-Madison [3] Univ. of Washington [4] Univ. of Toronto [5] Harvard Medical School [6] Massachusetts General Hospital [7] Williams College

We can machine learn anything…

- AI/ML is more **accurate and accessible** than ever
- **Appealing to predict** hard-to-measure outcomes
- **AI/ML-generated data** saturate fields of **genomics**, medicine, economics, demography, politics, etc.

Figure 1: Examples of papers conducting inference on an AI/ML generated outcome

… but then what?

- AI/ML-generated data often **reified as empirical**, raising questions of inferential validity
- Synthetic outcomes often **more correlated** with features of interest, **less variable**
- **Naïve use** in regressions leads to biased estimates, poor type 1 error control, and under-coverage

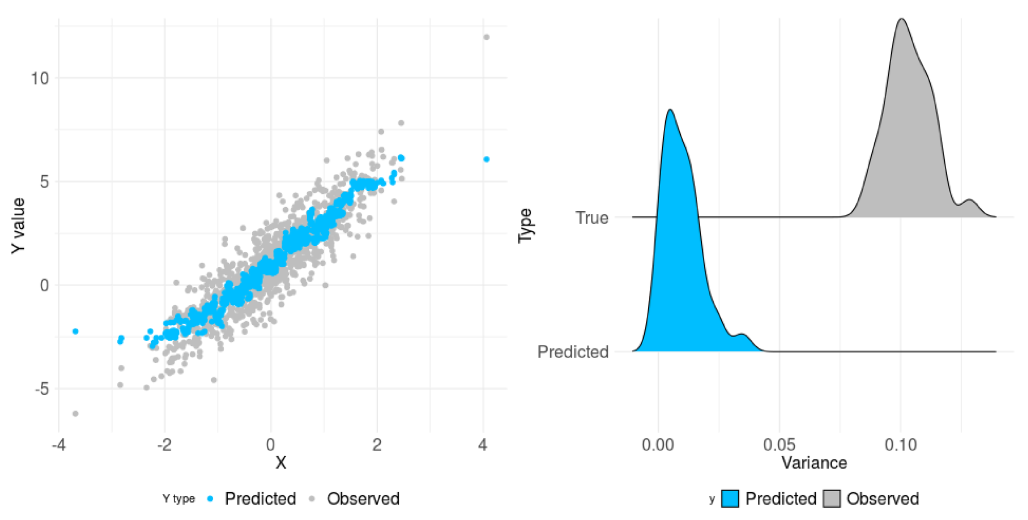
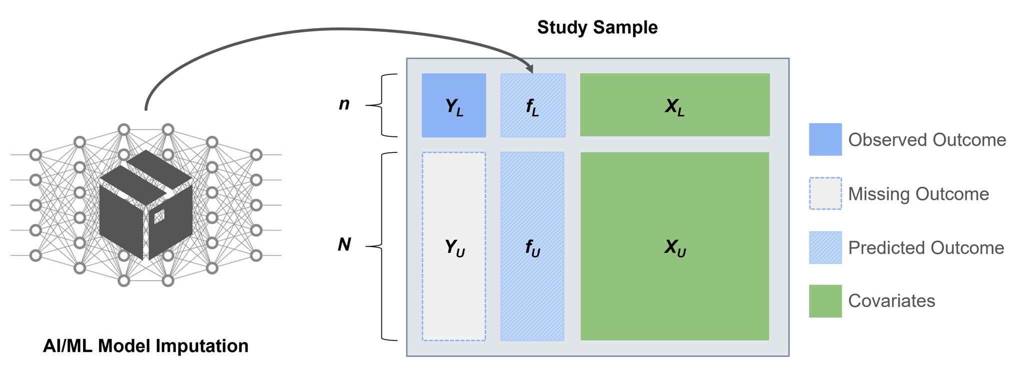


Figure 2: From Wang et al. (2020)

Inference with Predicted Data

Leverage some labeled data to **calibrate inference** in a study with mostly AI/ML-generated outcomes:

- $\boldsymbol{X}$: features, Y: true (partially observed) outcome
- $f : \mathcal{X} \to \mathcal{Y}$: prediction rule from training data, $f(\boldsymbol{X}) = \hat{Y}$: AI/ML-generated predictions
- Data: $\mathcal{L} = \{(X_i, Y_i)\}_{i=1}^{n} \cup \mathcal{U} = \{X_i\}_{i=n+1}^{n+N}$



Inference with Predicted Data (IPD) is a rapidly evolving field, driven by need for rigorous methods!

AI/ML-generated data exist **everywhere**.

High predictive **accuracy** ≠ valid for downstream **inference**.

There are now methods for conducting **inference with predicted data**.

Fred Hutch
Cancer Center

The ipd Package

Implements recent methods, data generation, and `tidy` helpers, for easy model fitting and inspection.

- Provides domain experts **user-friendly access** to these tools for use in their respective fields
- Enables data scientists developing new methods a means to **facilitate comparisons** and **contribute**
- Will be **continuously updated** to include more methods and functions

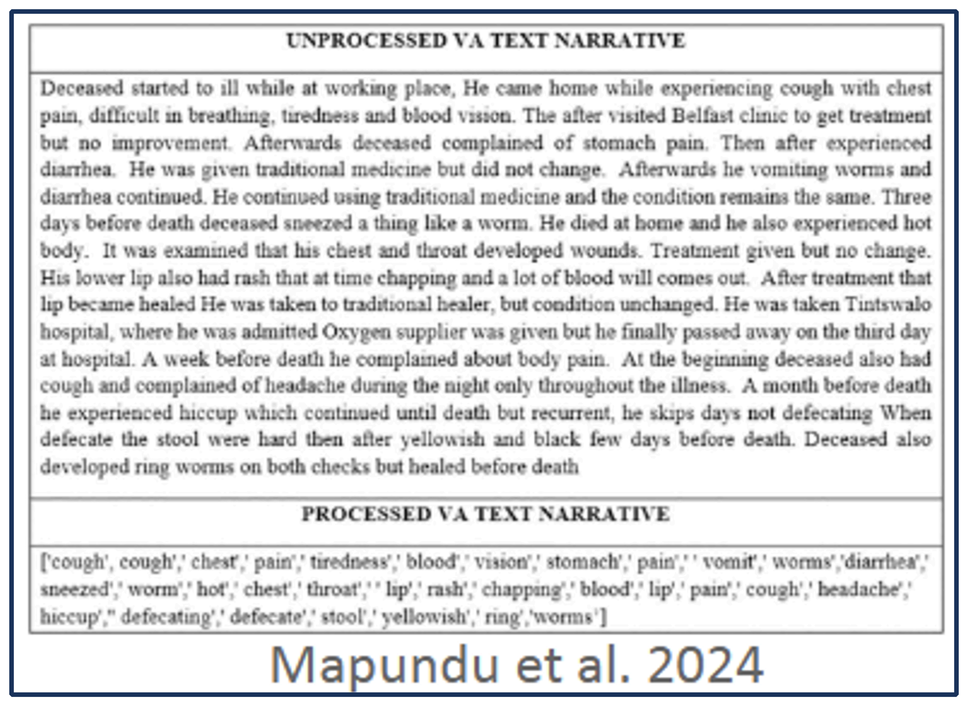
```
BiocManager::install("ipd")
library(ipd)
df <- simdat(model="ols") |>
  filter(set_label != "training")
fit <- ipd(Y - f ~ X1, method="chen", model="ols",
           data=df, label="set_label")
```

Open-source collaboration is the way to success!

Case Study: Verbal Autopsy (VA)

Context and Data:

- 2/3 of clinical **cause of death** (COD) certificates **missing** worldwide (Horton, 2007)
- VA involves **predicting** COD from **structured interviews** with family and caregivers
- The process is **time-consuming**, **resource-intensive**, and **error-prone**
- We have **gold-standard** labels and interviews for 6,763 deaths across 6 sites



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 vission. 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

Figure 3: Example VA narrative and tokenization (Mapundu et al., 2024)

Methods:

- VA interviews + **AI/ML** (KNN, Naïve Bayes, BERT, GPT-4) to **predict** 5 COD categories
- **Train** model in 5 sites, **estimate** in 6th
- Study **association** between age and COD

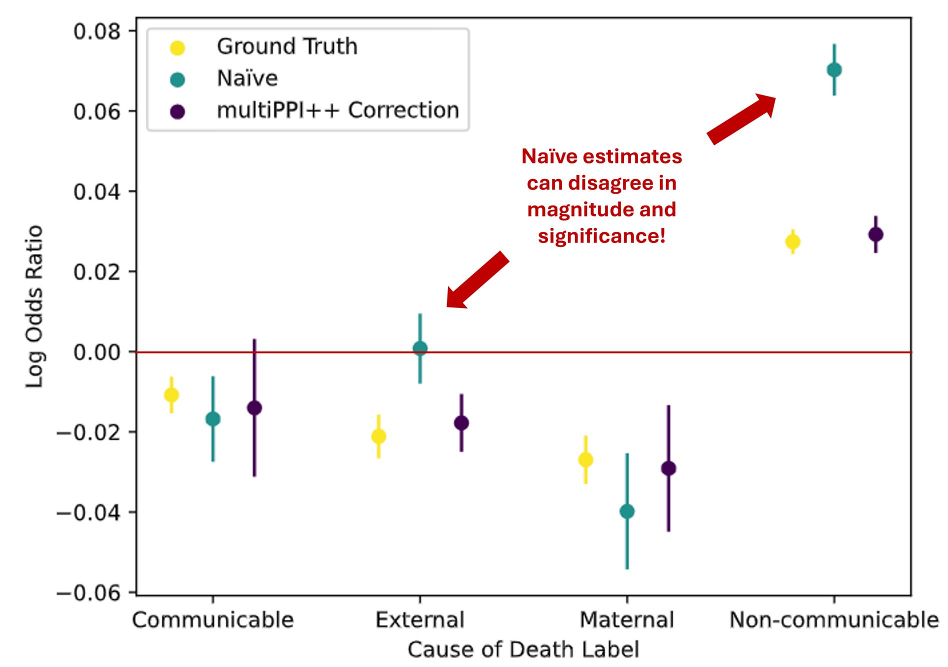


Figure 4: Example results from inference on Mexico site using KNN classifier (Fan et al., 2024)

Funding: NIH/NIGMS R35 GM144128 (SS, JTL), NIH/NHGRI U01 HG012039 (JM, QL), NIH/NIMH DP2 MH122405, R01 HD107015, and P2C HD042828 (THM).