AI, Data, & Implications for Demography

Tyler H McCormick
University of Washington
thmccormick.github.io





The Bureaucratic Regime

State-built legibility as a public good

Public agencies produce standardized statistics (and metadata) that organize people, firms, and places into comparable categories

Institutions

Census, BLS, BEA, NCES + OMB standards
Legal confidentiality; scheduled releases; public documentation
"Data modeling" → Testing theories about the world (spec. population groups)

Core values

Consistency → standardization → accountability
Same definitions over time enable trend claims & policy evaluation

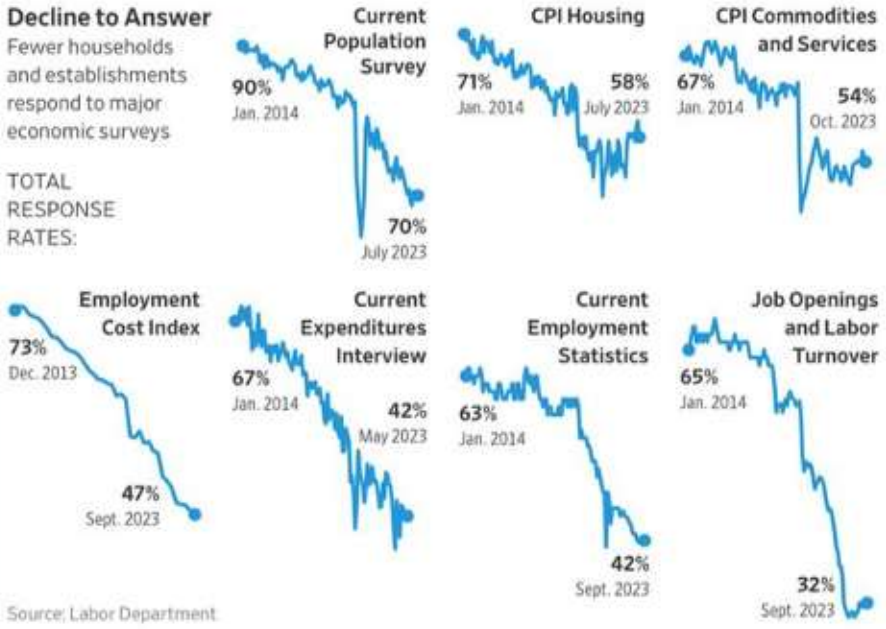
Examples

Decennial Census + ACS; CPS unemployment; CPI
NAEP; vital statistics; crime victimization surveys



Contradictions of the Bureaucratic Regime

- Survey response rates collapse
- **Cost**: Survey infrastructures are expensive to maintain → budget/staff cuts → loss of institutional knowledge and capacity
- **Distrust**: Acceleration of reputational attacks and integrity pressures (delayed releases, leadership churn, some data collection projects axed)
- **Competition**: Demands for transparency + data openness fuel the rise of private-sector data products and data analytics



Our new reality….

f(DataYouCanGet)=(DataYouWant)-ish

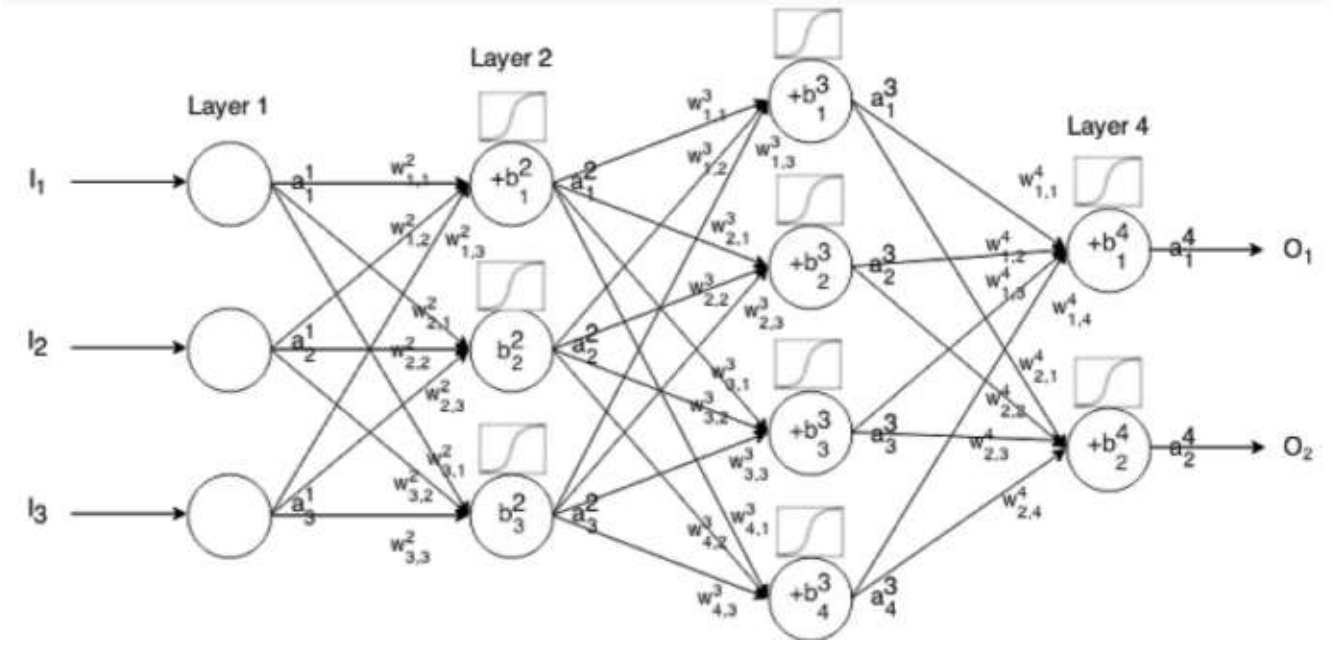


Image: https://www.cleanpng.com/png-artificial-neural-network-neuron-biological-neural-1616821/

Data collection and design

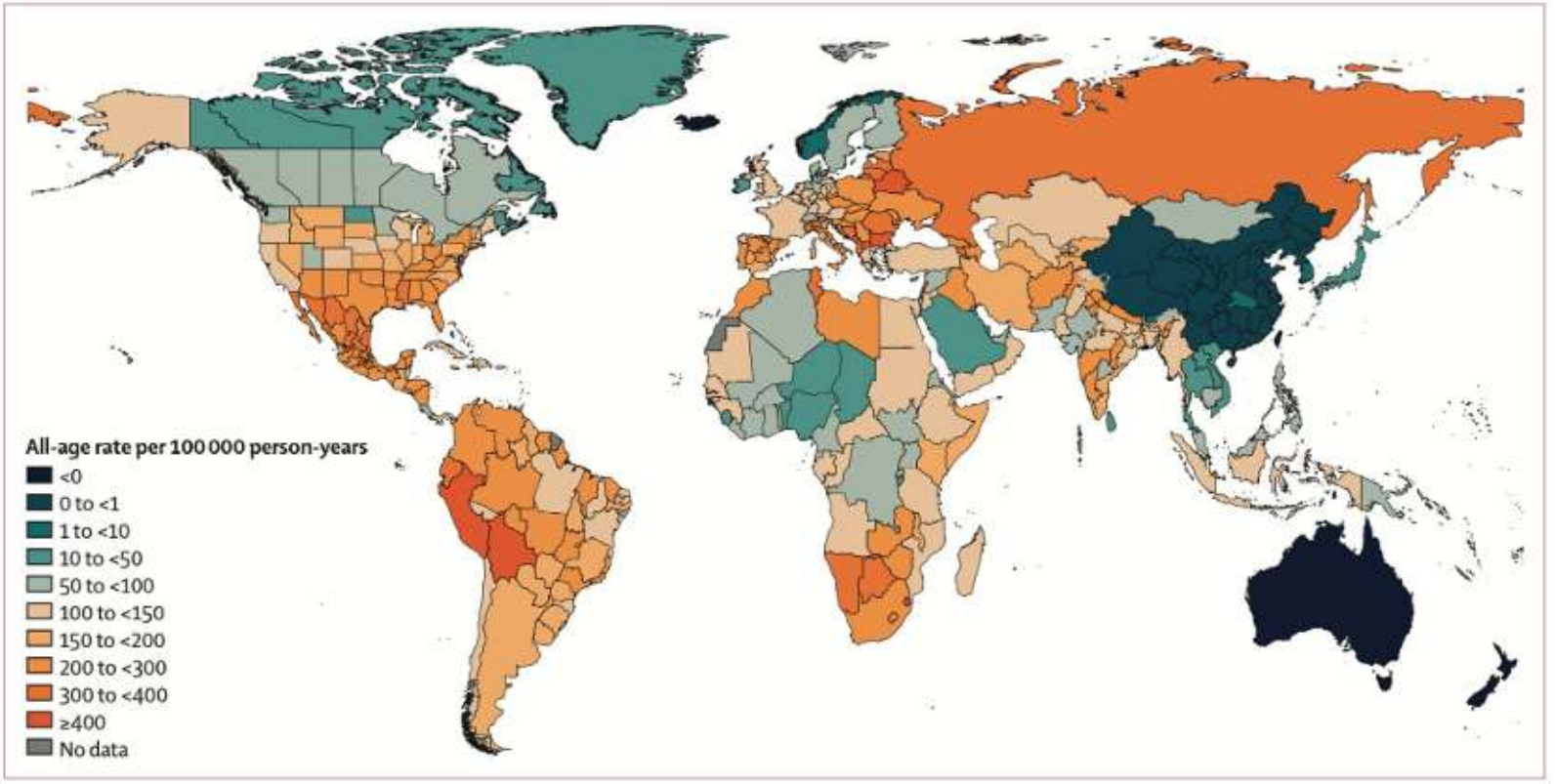
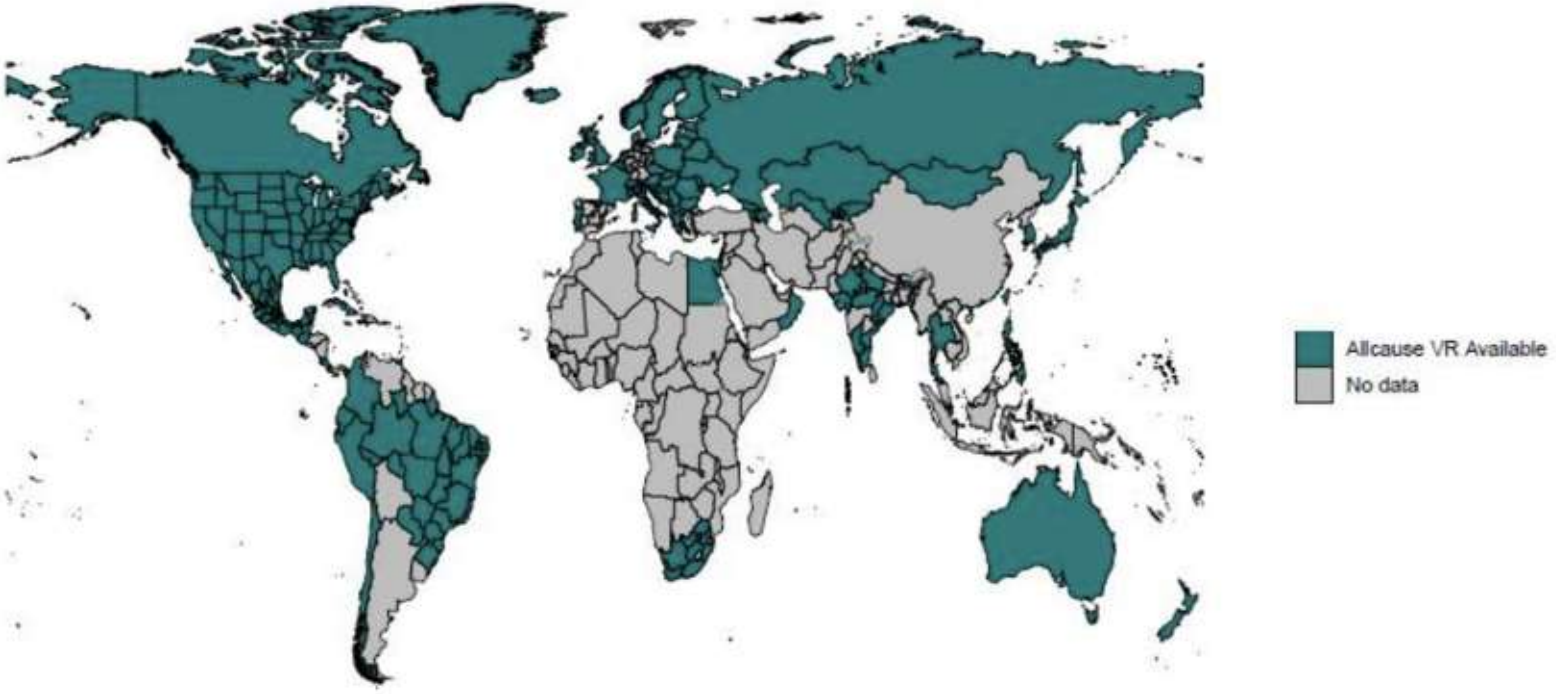


Figure 2: **Global distribution of estimated excess mortality rate due to the COVID-19 pandemic, for the cumulative period 2020–21**

IHME, March 2022, *The Lancet*

Data collection and design

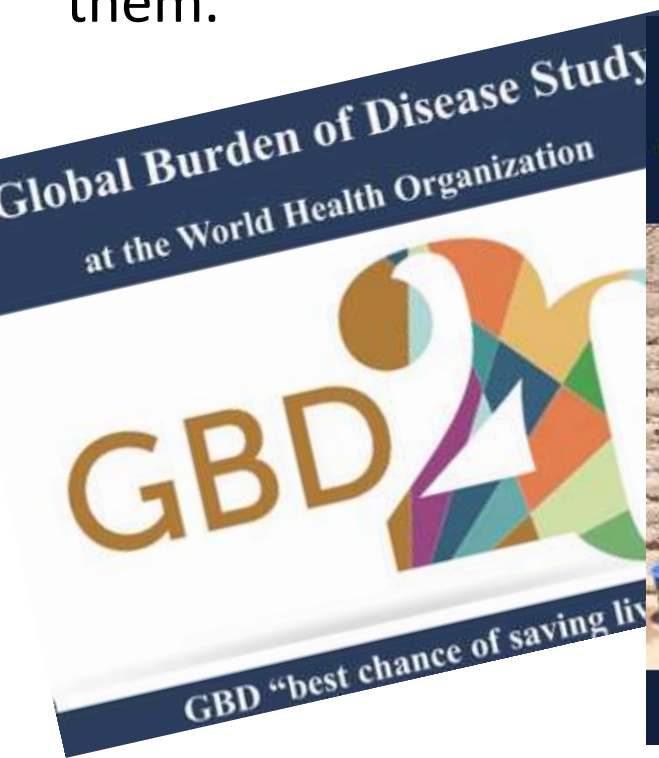
Figure S1. Map of all cause data availability



IHME, March 2022, *The Lancet*

Predictions are commodities

They have value and *institutions* exist to support and legitimize them.









The Brokerage Regime

State legitimated by pursuit of efficiency

Diversification of data and data producers. Anything (including administrative data) can be repurposed for prediction.

Institutions

State as purveyor of data to the private sector → "data products"
Government is one source of data among many
"Algorithmic modeling" → Detecting patterns and predicting outcomes, often for specific individuals

Core values

Speed ("nowcasting") + granularity + actionability
Algorithmic opacity + lock-in

Examples

Substitution of public data by private proxies
Merging of public data sources [no more "silos"] + augmentation of administrative data by commercial data [Palantir]
Public data as fuel for AI engines



Dilemmas of the Brokerage Regime

State governance after data decentralization

Statistical Rationality
Is predicted data "good enough"?

Political Rationality
Hollowing out: Parts of the state lose capacity to collect information and allocate resources toward public (democratic) goals

Authoritarian creep: Enhanced capacity for individual surveillance and targeting changes how the state thinks and acts

Political-Economic Restructuring: Displacement of public decision-making by private analytics reduces democratic accountability while turning various parts of the state into sources of private profit



Good prediction is not enough!!!!

Predictions are evaluated based on average performance in the test set

Regressions impacted by correlations in errors

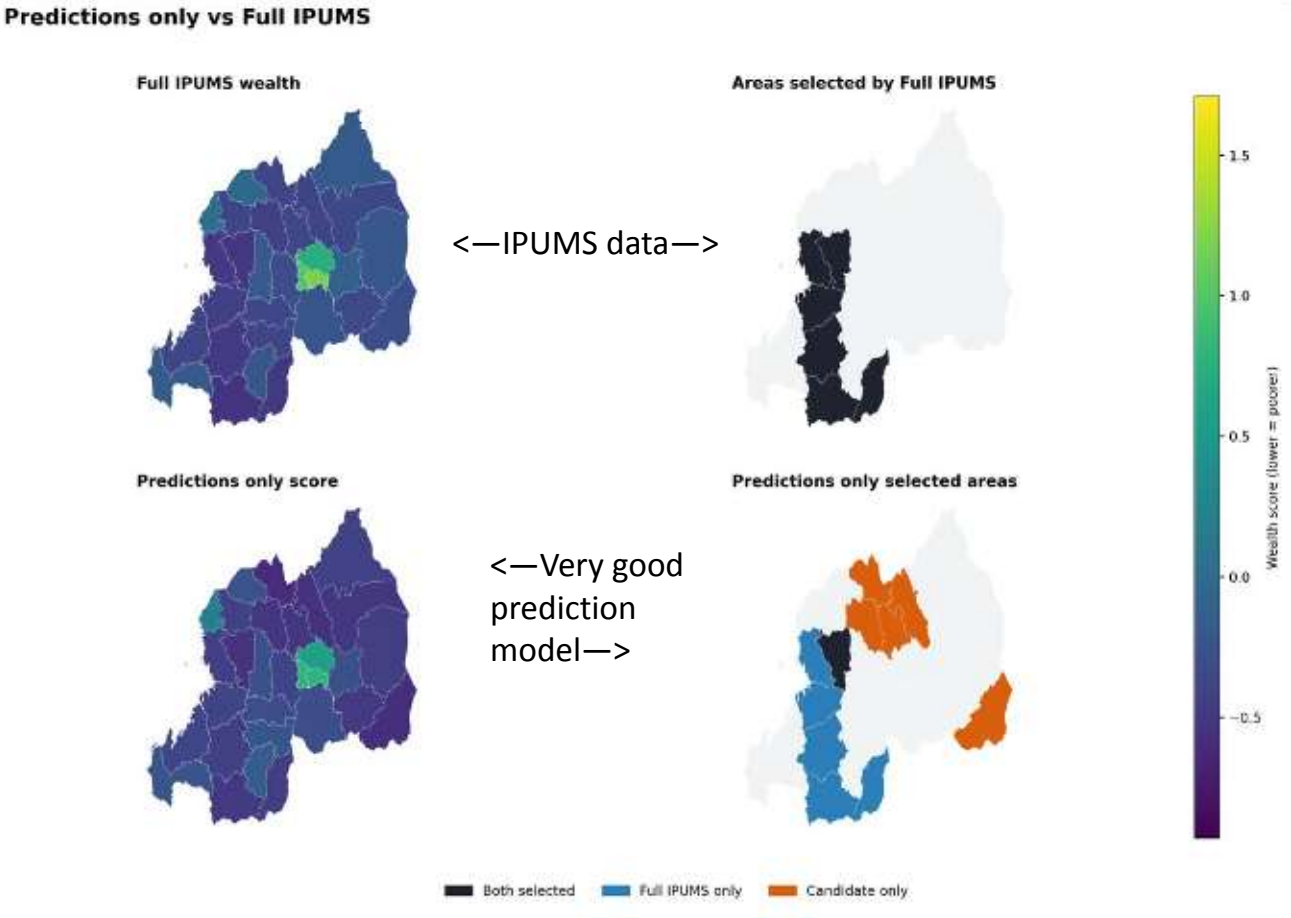
Policies depend on near local accuracy.

Rwanda, excellent predictor gives very different policies!

Using a prediction model for allocation selects regions that are not in the target wealth quintile and excludes the regions that are.

There's only one region in common between the two allocations strategies, despite the model having very high prediction accuracy.

Prediction model: Yeh et al, 2020

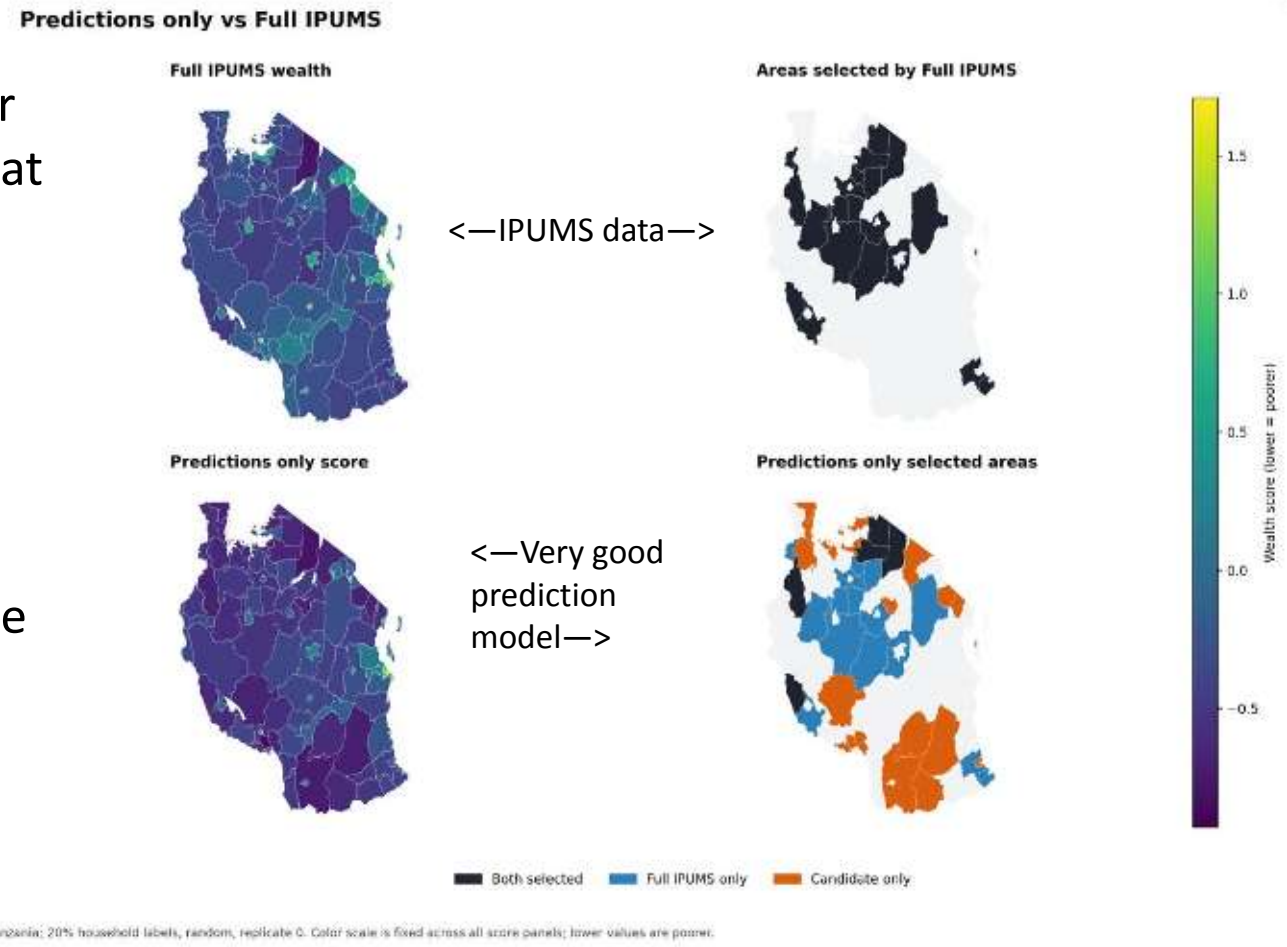


Rwanda, 70% household labels, random, replicate 0. Color scale is fixed across all score panels; lower values are poorer.

Tanzania, again very different policies!

Using a prediction model for allocation selects regions that are not in the target wealth quintile and excludes the regions that are.

Only 5 of the 15 regions in desired wealth quintile are selected for allocation by the prediction model.



Prediction model: Yeh et al, 2020

Only one solution....



Statistical calibration

SOME FURTHER RESULTS ON ERRORS IN DOUBLE SAMPLING TECHNIQUE*

By CHAMELI BOSE
Statistical Laboratory, Calcutta

1. The author (1942, 1943) studied some types of double sampling technique which consists in obtaining an expression for a character y, sometimes difficult or uneconomic to measure directly, in terms of an appreciably correlated character x, easier to obtain. Such problems were studied also by Neyman (1938), and Cochran (1939); but their methods of approach were different. Snedecor and King (1942) obtained a formula for the variance of the forecasting equation for

Sampling Techniques

third edition

1977

WILLIAM G. COCHRAN

Professor of Statistics, Emeritus
Harvard University

12.6 REGRESSION ESTIMATORS

In some applications of double sampling the auxiliary variate x_i has been used to make a regression estimate of $\bar{Y}$. In the first (large) sample of size n', we measure only x_i; in the second, a random subsample of size $n = \nu n' = n'/k$ where the fraction ν is chosen in advance, we measure both x_i and y_i. The estimate of $\bar{Y}$ is

$$\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x}) \tag{12.48}$$

where $\bar{x}'$, $\bar{x}$ are the means of the x_i in the first and second samples and b is the least squares regression coefficient of y_i on x_i, computed from the second sample.

Summary: challenges and opportunities

- Design and data collection are more important than ever!
  - Opportunistic data collection
  - Representativeness; who/what are we excluding?
  - True in RCTs and learning associations
- Prioritize learning and be strategic (this idea is not new!)

Good performance on prediction is not enough!!!

Statistical calibration can work extremely well, but relies on having targeted ***validation data***!

Link to these slides



thmccormick.github.io

tylermc@uw.edu

Link to paper on data regimes

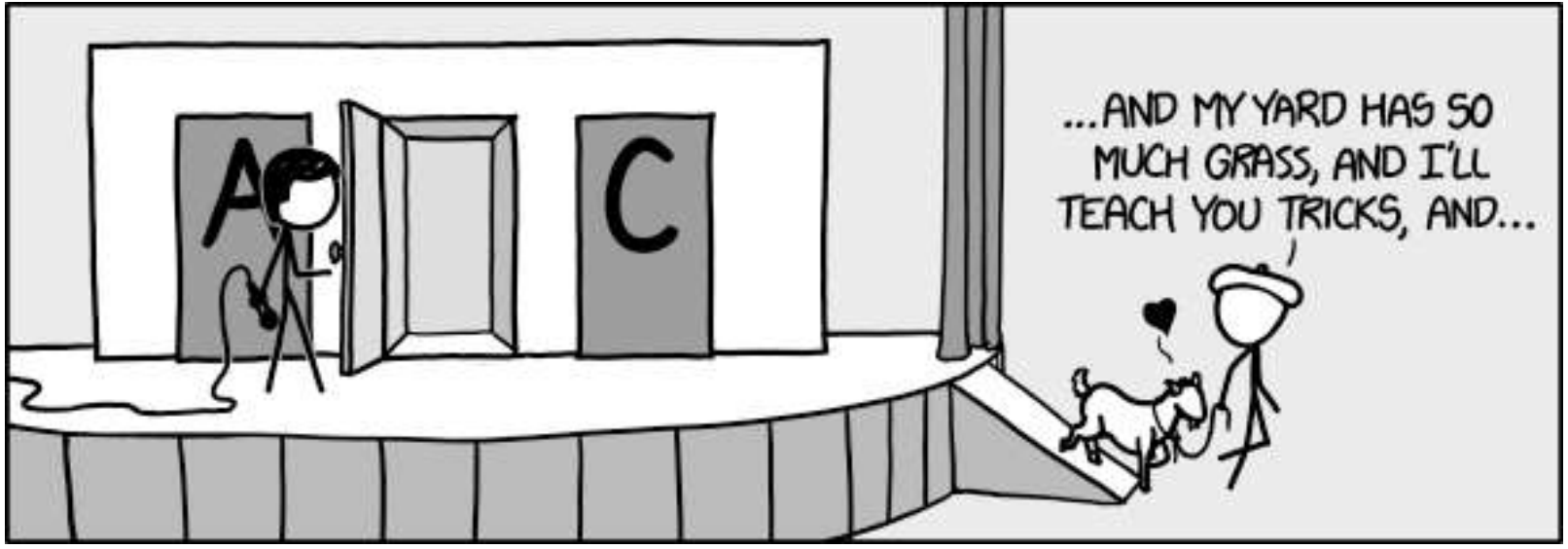


thmccormick.github.io

tylermc@uw.edu

Thank you!

tylermc@uw.edu
thmccormick.github.io



https://imgs.xkcd.com/comics/monty_hall.png

Further reading

- Meng, Xiao-Li. "Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election." *The Annals of Applied Statistics* 12.2 (2018): 685-726.
- Yu, Bin. "Veridical data science." *Proceedings of the 13th international conference on web search and data mining*. 2020.
- Rudin, Cynthia, et al. "Amazing things come from having many good models." *arXiv preprint arXiv:2407.04846* (2024).
- Parikh, Harsh, et al. "Who Are We Missing?: A Principled Approach to Characterizing the Underrepresented Population." *Journal of the American Statistical Association* just-accepted (2025): 1-32.
- Wang, Siruo, Tyler H. McCormick, and Jeffrey T. Leek. "Methods for correcting inference based on outcomes predicted by machine learning." *Proceedings of the National Academy of Sciences* 117.48 (2020): 30266-30275.
- Angelopoulos, Anastasios N., et al. "Prediction-powered inference." *Science* 382.6671 (2023): 669-674.
- Ji, Wenlong, Lihua Lei, and Tijana Zrnic. "Predictions as surrogates: Revisiting surrogate outcomes in the age of ai." *arXiv preprint arXiv:2501.09731* (2025).
- Gronsbell, Jessica, et al. "Another look at inference after prediction." *arXiv preprint arXiv:2411.19908* (2024).
- Miao, Jiacheng, et al. "Assumption-lean and data-adaptive post-prediction inference." *arXiv preprint arXiv:2311.14220* (2023).

ipd: Inference on Predicted Data



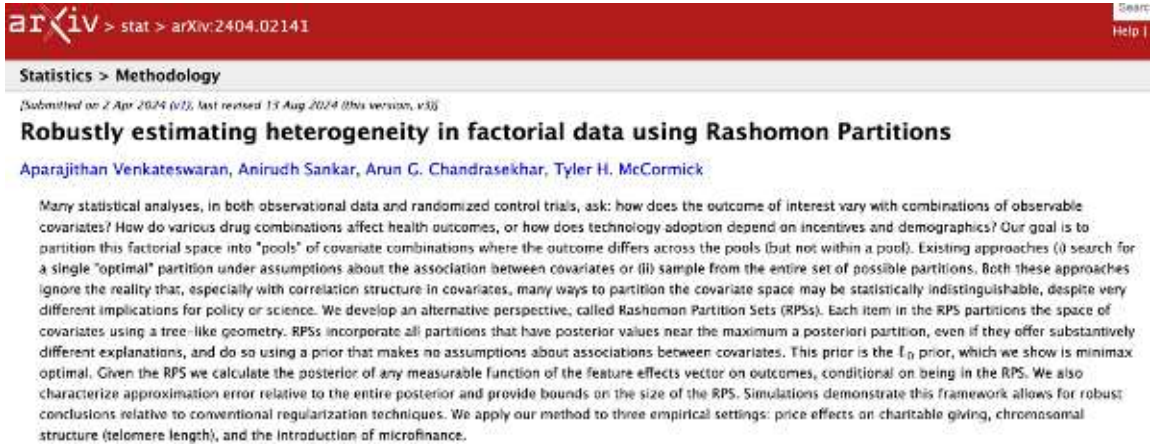
Overview

`ipd` is an open-source `R` software package for the downstream modeling of an outcome and its associated features where a potentially sizable portion of the outcome data has been imputed by an artificial intelligence or machine learning (AI/ML) prediction algorithm. The package implements several recent proposed methods for inference on predicted data (IPD) with a single, user-friendly wrapper function, `ipd`. The package also provides custom `print`, `summary`, `tidy`, `glance`, and `augment` methods to facilitate easy model inspection.

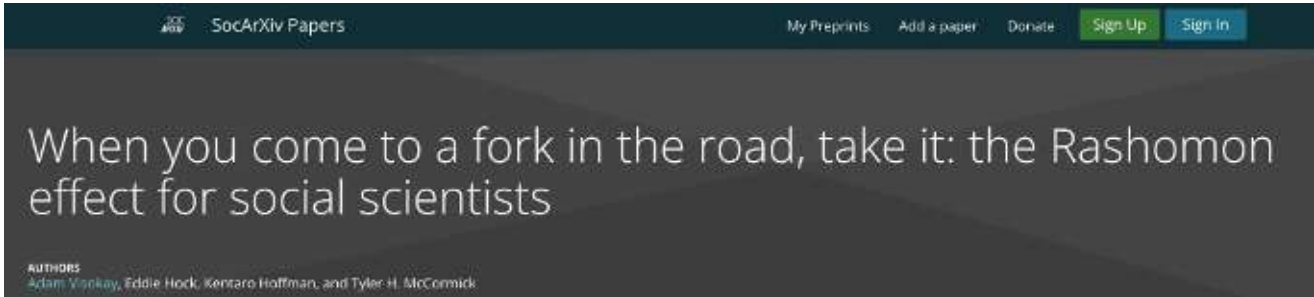


Our work on Rashomon sets









Our work on predictions and inference

BMI



Verbal autopsy









A “given data” approach

Based on how we teach statistics...

Suppose that $X_1, \ldots, X_n$ are independent and identica
from a $N(\mu_1, \sigma^2)$ distribution, where μ_1 and σ^2 are unk

Suppose also that $Y_1, \ldots, Y_m$ are independent
from a $N(\mu_2, \sigma^2)$ distribution

Find the max

Suppose that a coin is tossed repeatedly. Each toss is independent. with probability p of heads. Let X represent the number of tails befor the first heads. Let $\alpha = 0.05$.

(a) Suppose we are interested in testing $H_0 : p = 1/2$ versus $H_1 : p$ p_1 for some fixed $p_1 < 1/2$. Determine the most powerful leve test.

(b) Suppose we are interested in testing $H_0 : p = 1/2$ versus $H_1 : p$ $1/2$. Determine whether the test from (a) is the uniformly mos

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed with $X_i \sim N(0,1)$. Furthermore, for $i = 1, \ldots, n$, let $Y_i \mid X_i = x \sim N(\theta x, 1)$.

(a) Find the log likelihood function corresponding to $(X_1, Y_1), \ldots, (X_n, Y_n)$.

(b) Find the maximum likelihood estimator $\hat{\theta}_n$ of θ.

Let $(Y_1, X_1), \ldots, (Y_n, X_n) \in \mathbb{R}^2$ be independent and ide tributed random vectors from an unknown distribution, i (Y_1, X_1) is independent of the pair (Y_2, X_2), but Y_1 may b on X_1. Furthermore, both X_i and Y_i are continuous wi probability density functions. We would like to estimate μ

However, we do not observe Y_i. Instead, we observe

$$(Z_1, X_1), \ldots, (Z_n, X_n),$$

Let $X_1, \ldots, X_n$ be inde ables with probability density fun

$$f_X(x;\theta) = \frac{1}{2}(1 + \theta x) I_{[-1,1]}(x),$$

with θ unknown, $0 \le \theta \le 1$. The notation $I_{[-1,1]}(x)$ repre indicator variable that equals 1 if $-1 \le x \le 1$, and 0 otherwi

(a) Use the method of moments to find an estimator $\hat{\theta}_n$ of

(b) Is $\hat{\theta}_n$ consistent?

(c) Find the asymptotic distribution of $\hat{\theta}_n$.

(d) Find a level $(1 - \alpha)$ confidence interval for θ.

TIdr;

Predictions are ***evaluated*** based on matching answers in the test set, but ***used*** for statistical inference or policy decision-making.

Data collection and design

Tldr; You can trade on predictions, but no substitute for data.



A big caveat!!

Statistical calibration can work extremely well, but relies on having targeted ***validation data***!

A data revolution?

Many problems in demography involve using the data you can easily access to ***predict*** the variables you want.

Data not designed for research

Bureaucratic regime⟶ brokerage regime

TIdr;

Predictions are evaluated based on average performance in the test set

Regressions impacted by correlations in errors

Policies depend on near local accuracy.



SSRN

Product & Services Subscribe Submit a paper Browse More

Download This Paper Open PDF in Browser

Add Paper to My Library

From Census to Sensors: The Bureaucratic Regime and the Data Revolution

49 Pages • Posted: 12 Jun 2026

Marion Fourcade
University of California, Berkeley - Department of Sociology

Tyler McCormick
University of Washington

Adam Visokay
University of Washington

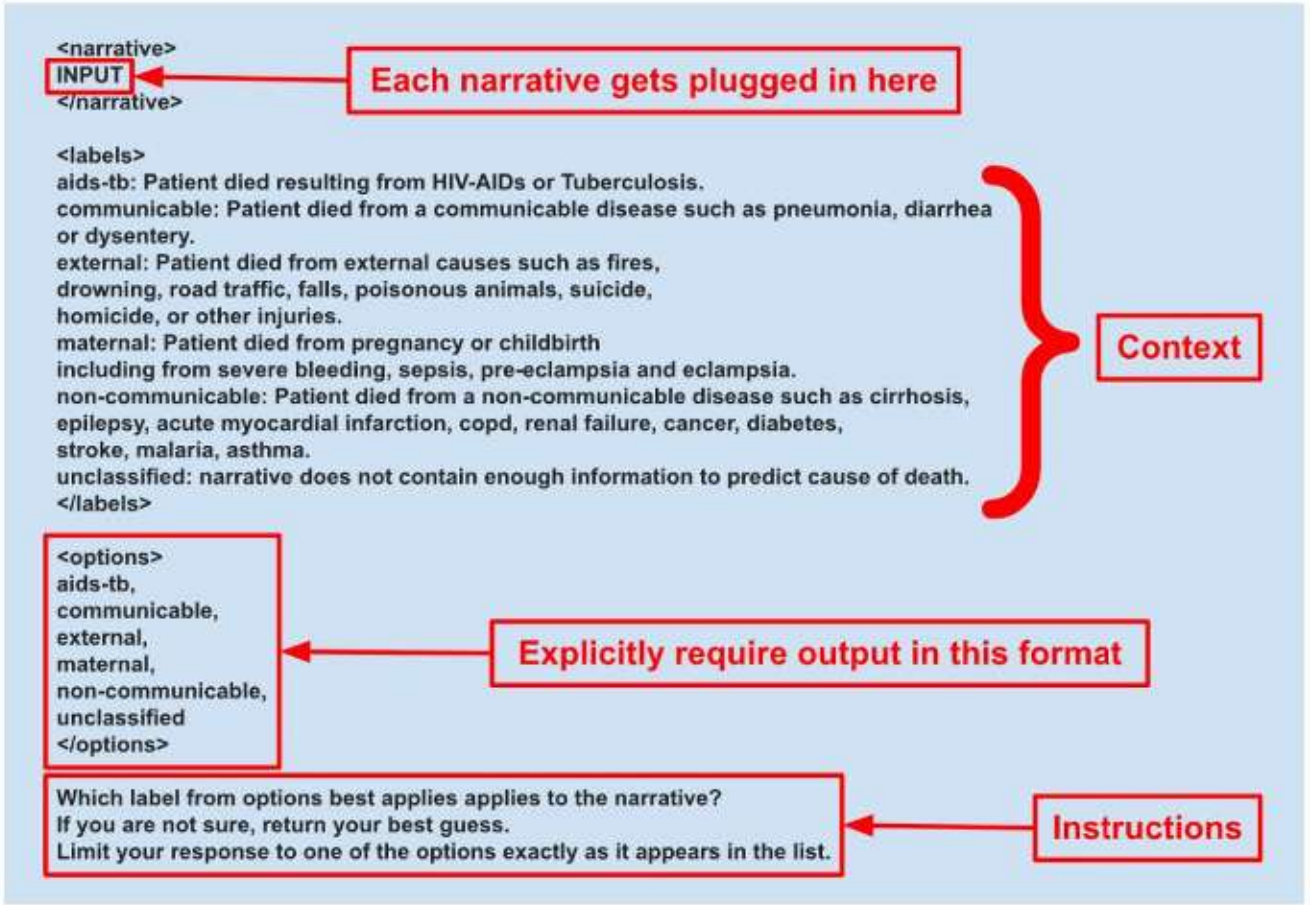
Sarah Quinn
University of Washington

Date Written: May 15, 2026

Example 1: COD prediction w/ zero-shot GPT

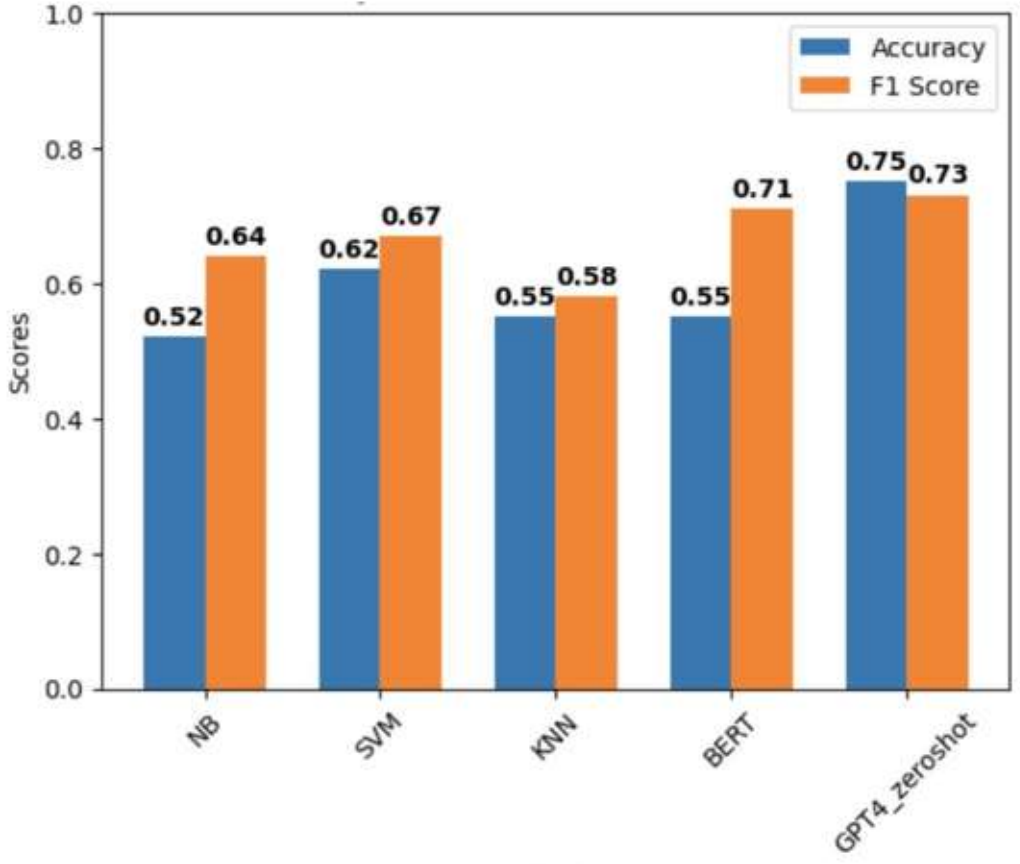
Deaths happen outside of hospital

Recordings of description of circumstances from caregiver/family member



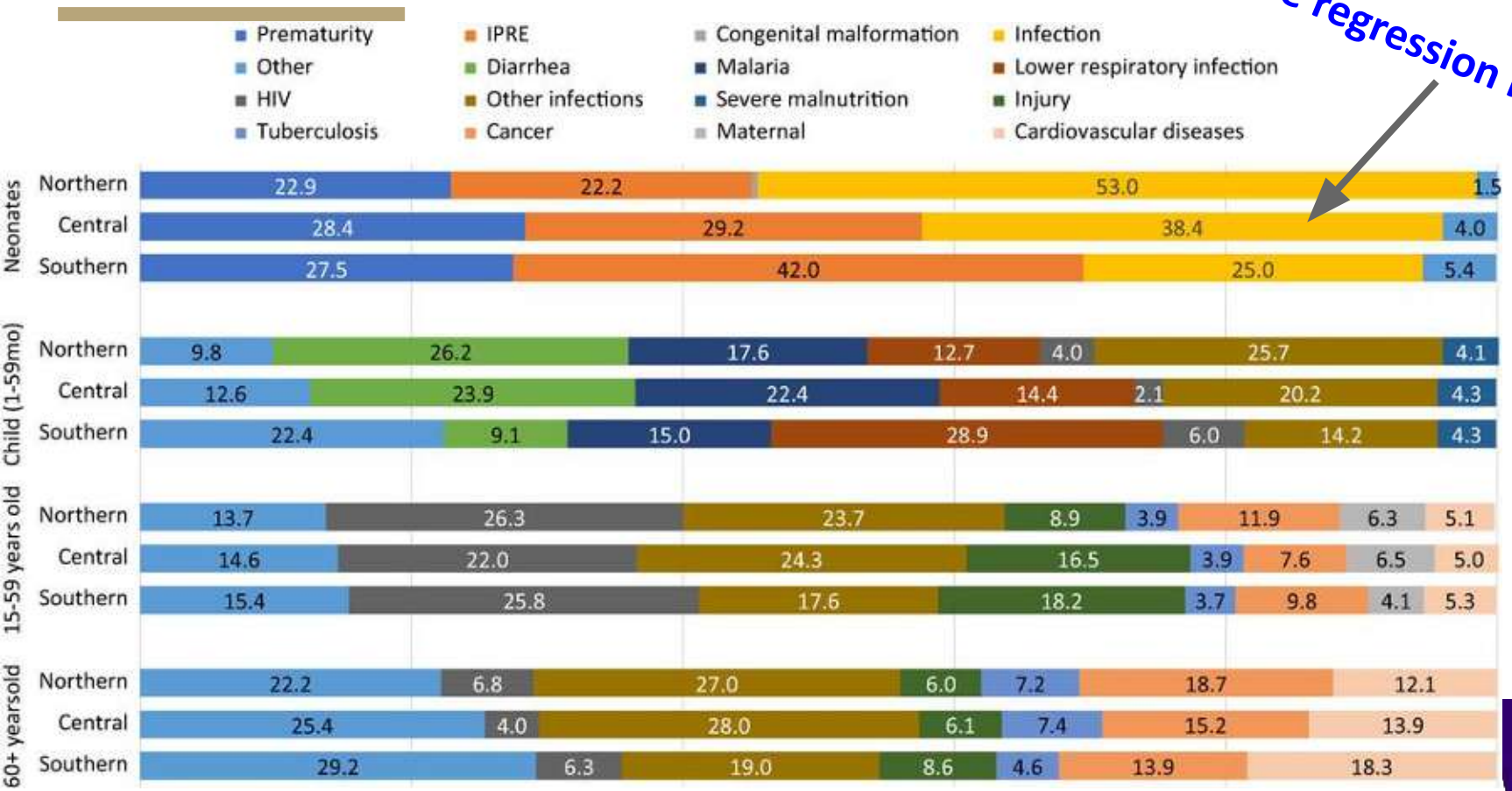
Results: classification

F1 score “balances”
false negs and false pos



Statistical questions (COD in Mozambique)

These are regression models!



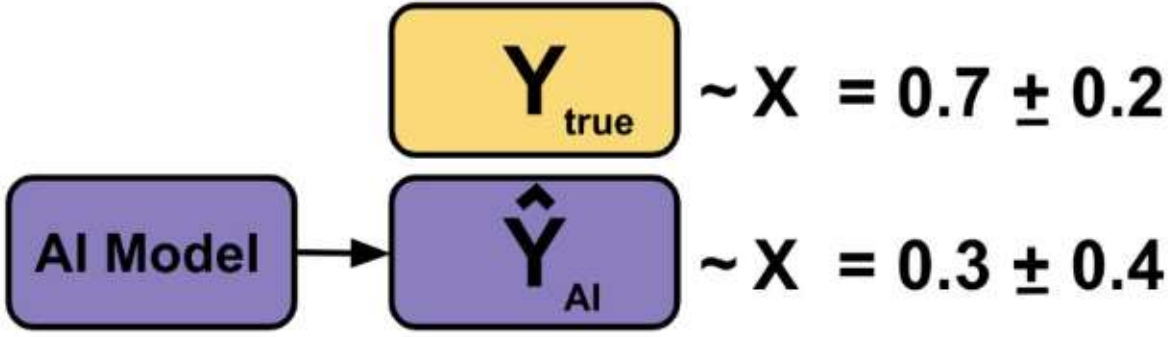
Source: Macicame *et al*, 2023, Am J Trop Med Hyg

What are we really doing?

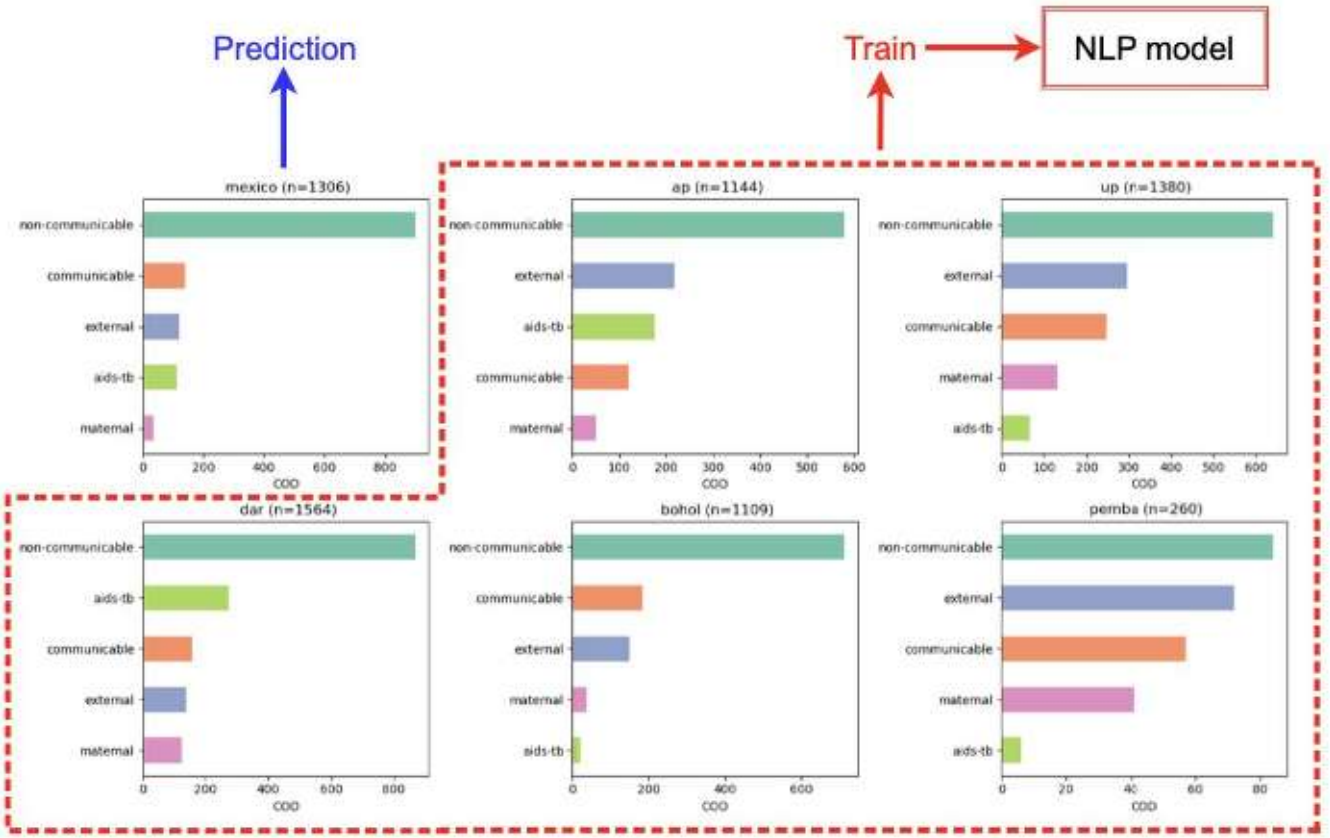
We're doing *inference* (regressions) using *predicted* outcomes (COD)

$$g(COD) = \beta_1 \text{age} + \beta_2 \text{region} + \epsilon$$

Without correction, we get invalid uncertainty *and* potential bias.

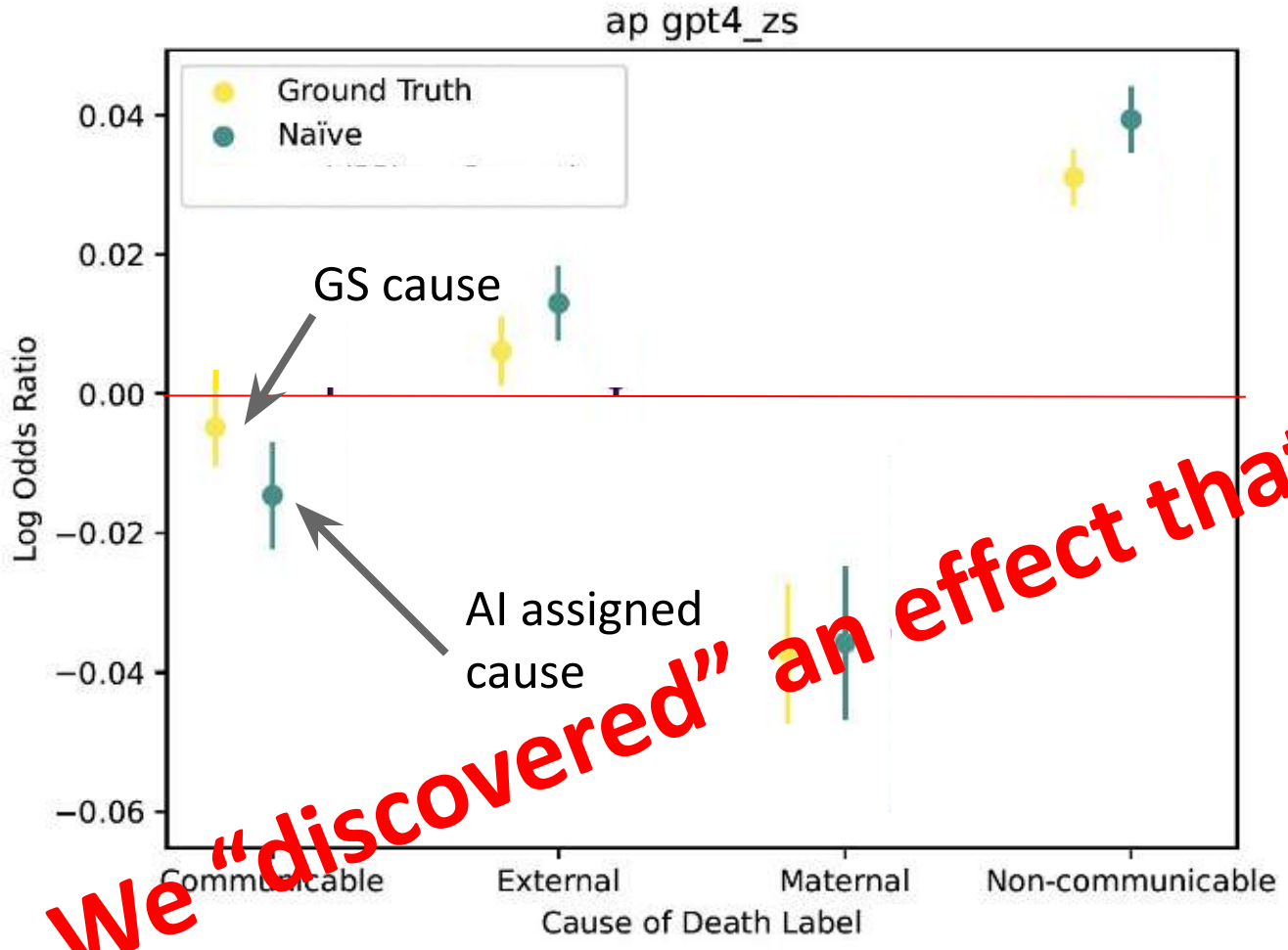


Train the NLP model with one site left out



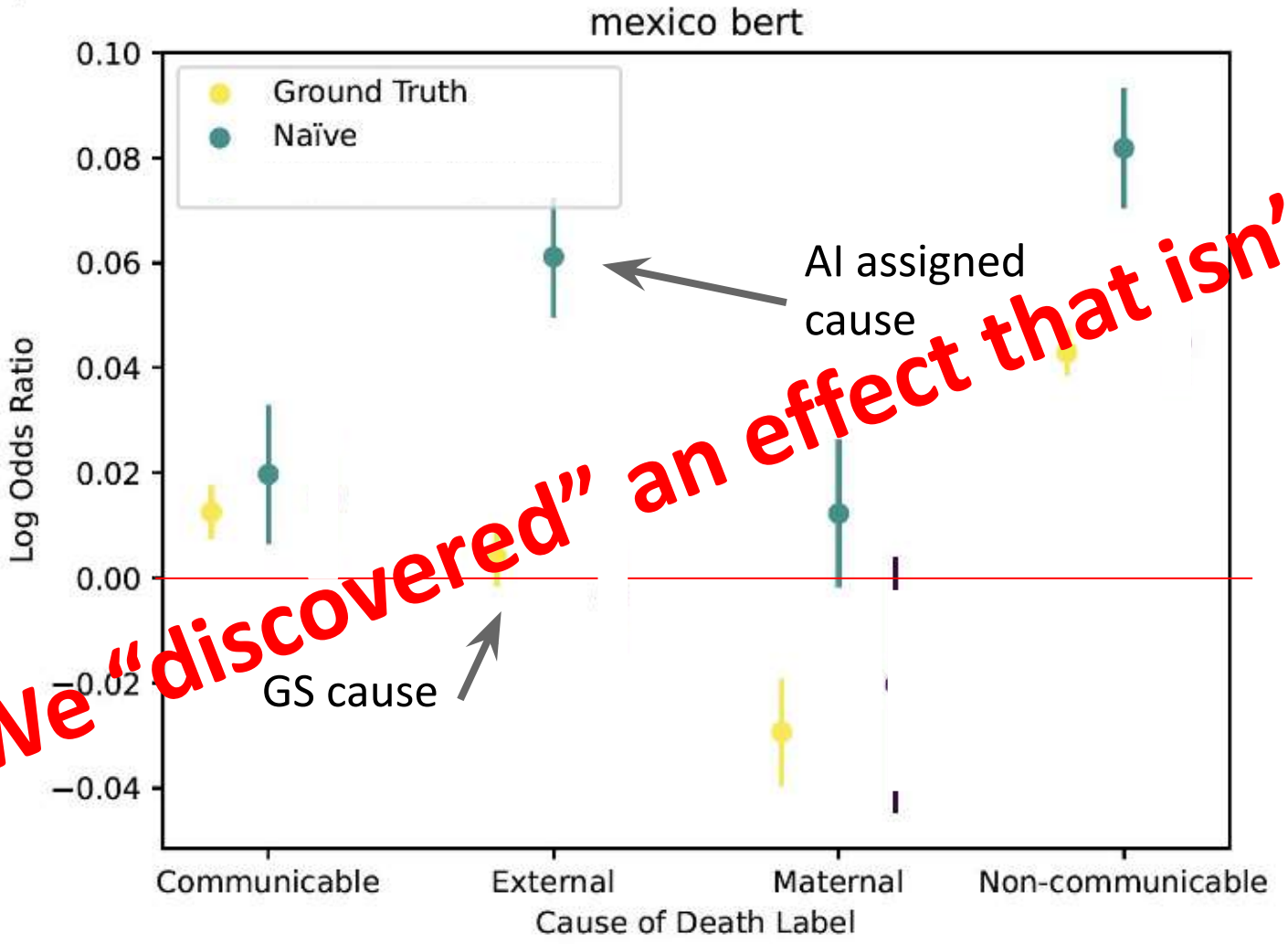
We use narratives to make predictions and both classic ML and modern AI techniques.

PHMRC Gold Standard Reference Data



g(cause)=beta0+ beta1*age+epsilon

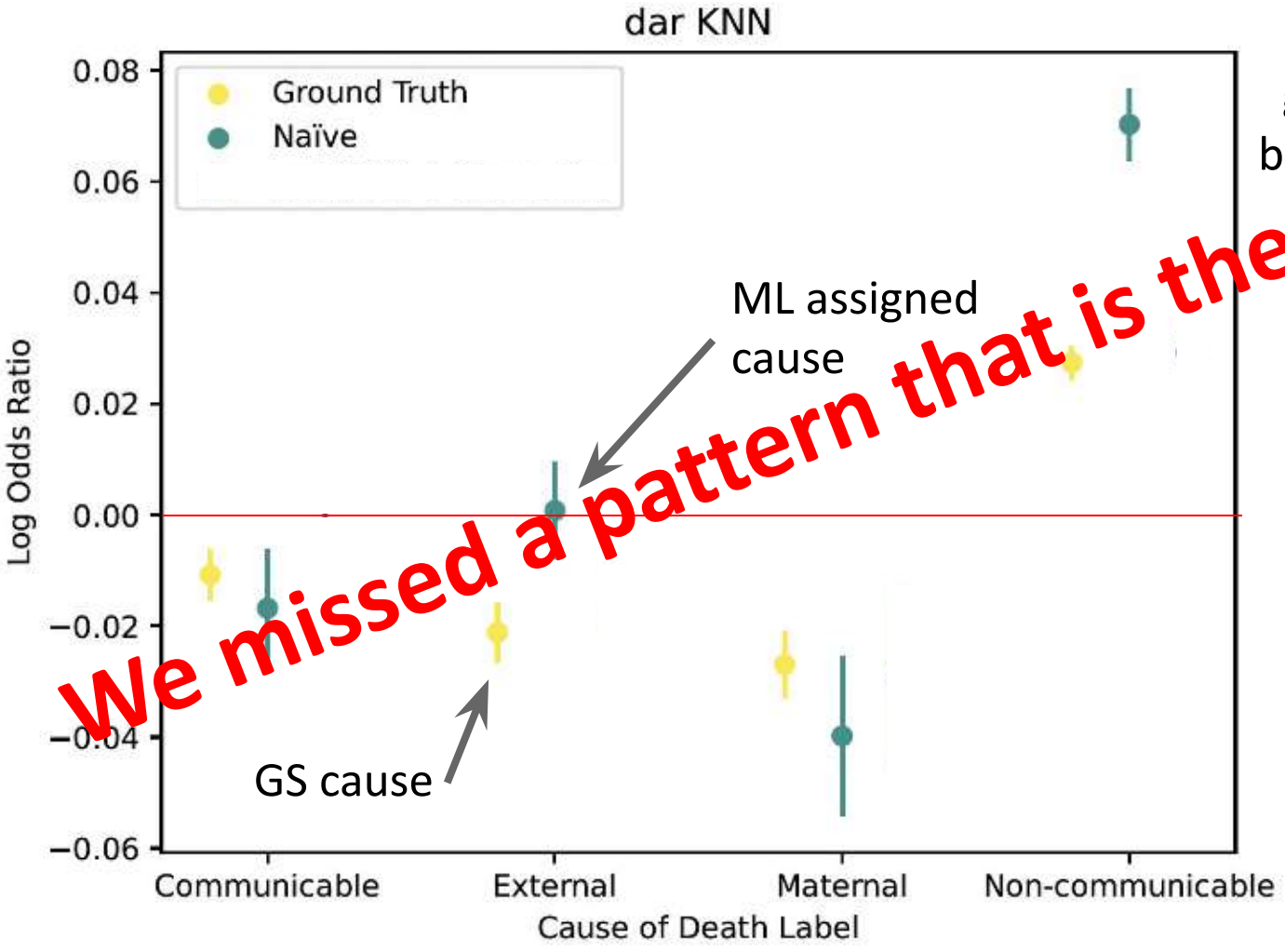
We "discovered" an effect that isn't there!!!



We "discovered" an effect that isn't there!!!!

Conclusions change dramatically!!!

This time with BERT.



g(cause)=beta0+ beta1*age+epsilon

Example 2: Geographic policy allocation

> Nat Commun. 2020 May 22;11(1):2583. doi: 10.1038/s41467-020-16185-w.

Using publicly available satellite imagery and deep learning to understand economic well-being in Africa

Christopher Yeh [# 1], Anthony Perez [# 1 2], Anne Driscoll [3], George Azzari [2 4], Zhongyi Tang [5], David Lobell [3 4 5], Stefano Ermon [1], Marshall Burke [6 7 8 9]

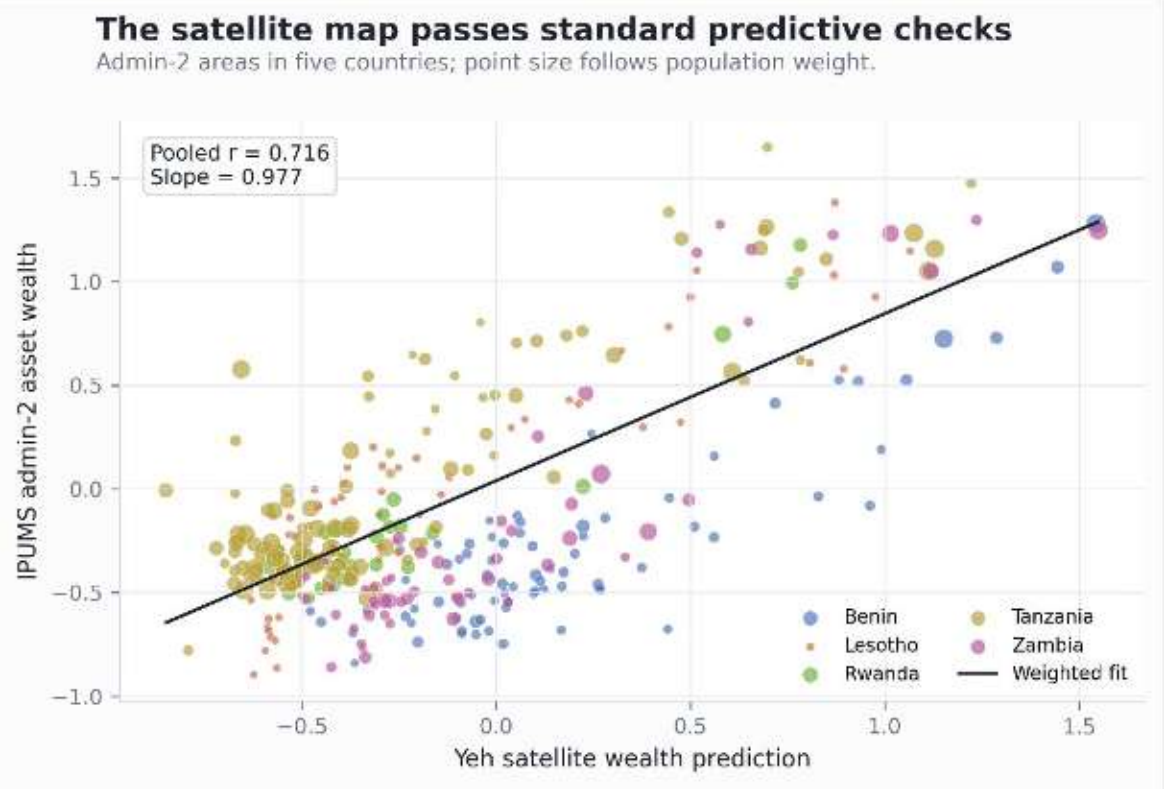
Affiliations + expand

PMID: 32444658 PMCID: PMC7244551 DOI: 10.1038/s41467-020-16185-w

Abstract

Accurate and comprehensive measurements of economic well-being are fundamental inputs into both research and policy, but such measures are unavailable at a local level in many parts of the

Comparison with IPUMS HH data in 5 countries



0.716

pooled admin-2 correlation with Full IPUMS wealth

0.861-0.950

country-level correlations

Geographic targeting

1

Define target

Residents of admin-2 areas in the within-country Full IPUMS bottom quintile.

2

Choose areas

Select about the bottom 20% of residents within country using an area score.

3

Reveal local labels

Sample 0.025% to 20% of households; attach Yeh score and pre-label covariates.

4

Evaluate policy quantities

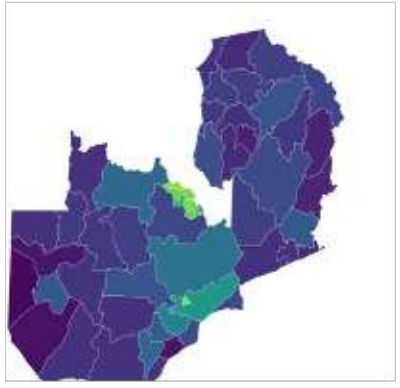
Mean wealth reached, bottom-quintile capture, undercoverage, leakage, overlap.

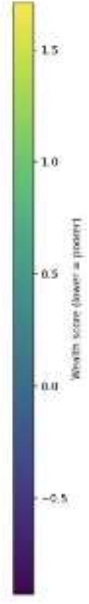
Common targeting strategy (e.g., IRS opportunity zones; HUD comm dev block grants)

Zambia, very different policies!

Black = both selected; blue = Full IPUMS only; orange = prediction rule only.

Full IPUMS wealth



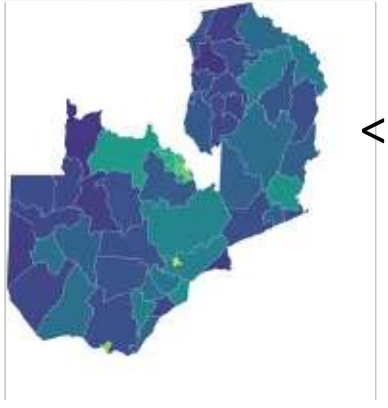


<—Full data—>

Areas selected by Full IPUMS



Predictions only score



<—Very good prediction model—>

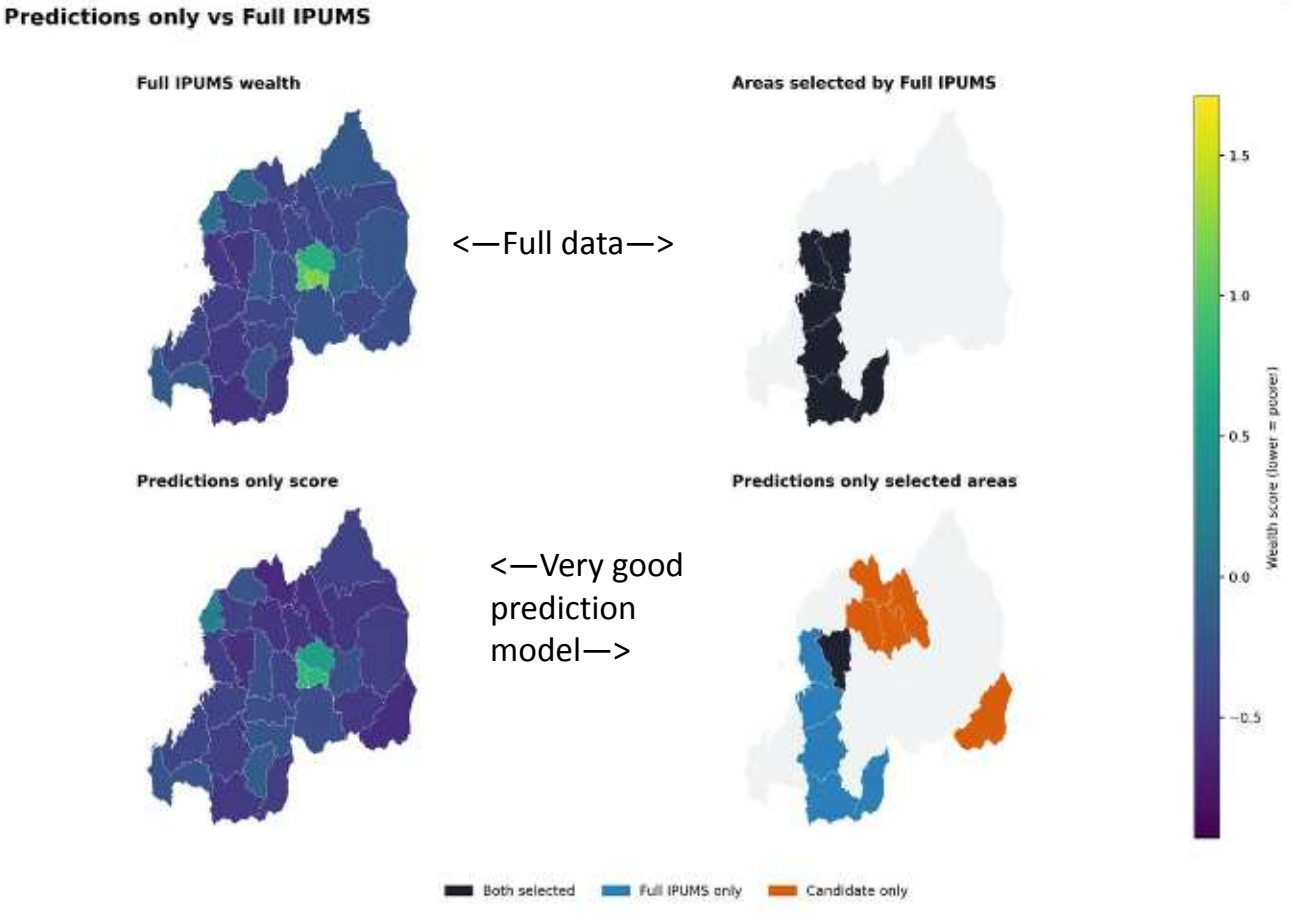
Predictions only selected areas



Rwanda, very different policies!

Using a prediction model for allocation selects regions that are not in the target wealth quintile and excludes the regions that are.

There's only one region in common between the two allocations strategies, despite the model having very high prediction accuracy.

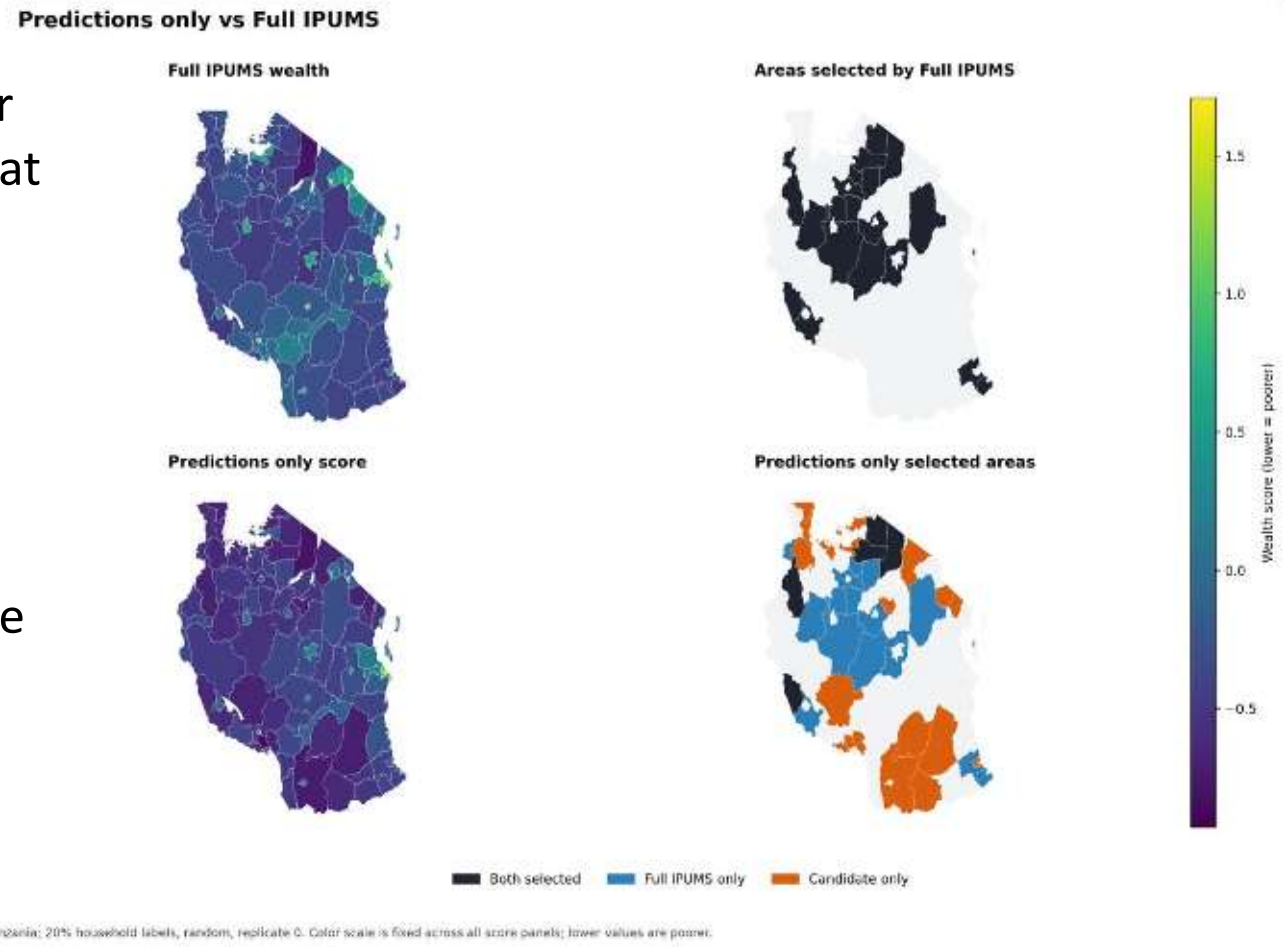


Rwanda, 20% household labels, random, replicate 0. Color scale is fixed across all score panels; lower values are poorer.

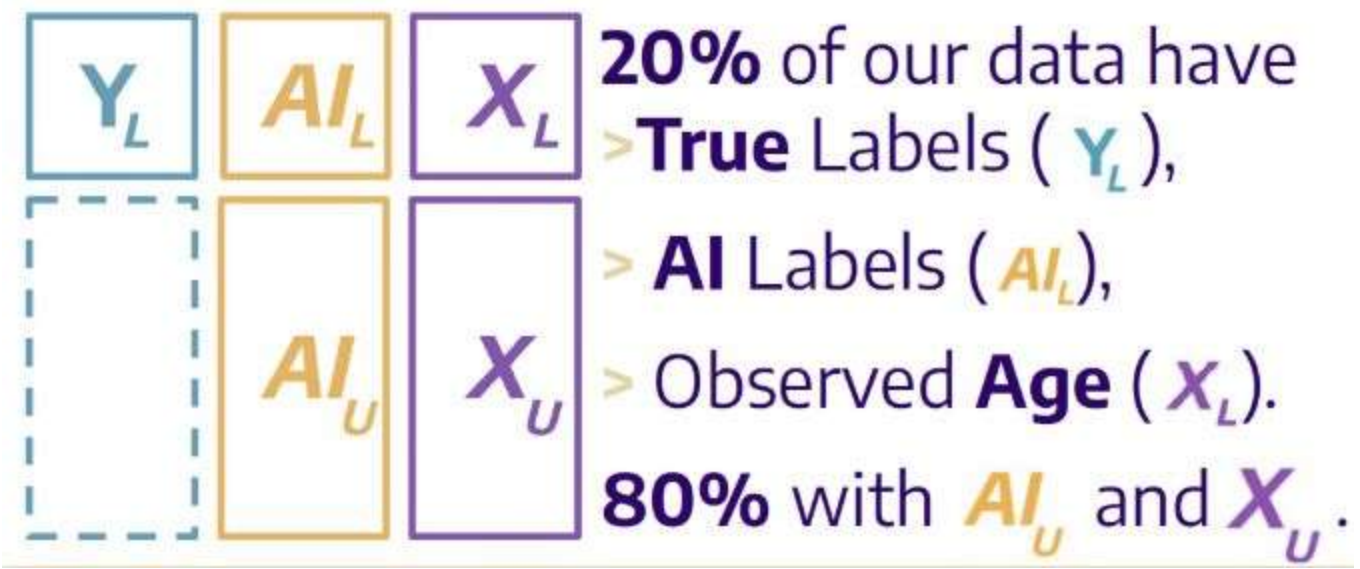
Tanzania, very different policies!

Using a prediction model for allocation selects regions that are not in the target wealth quintile and excludes the regions that are.

Only 5 of the 15 regions in desired wealth quintile are selected for allocation by the prediction model.

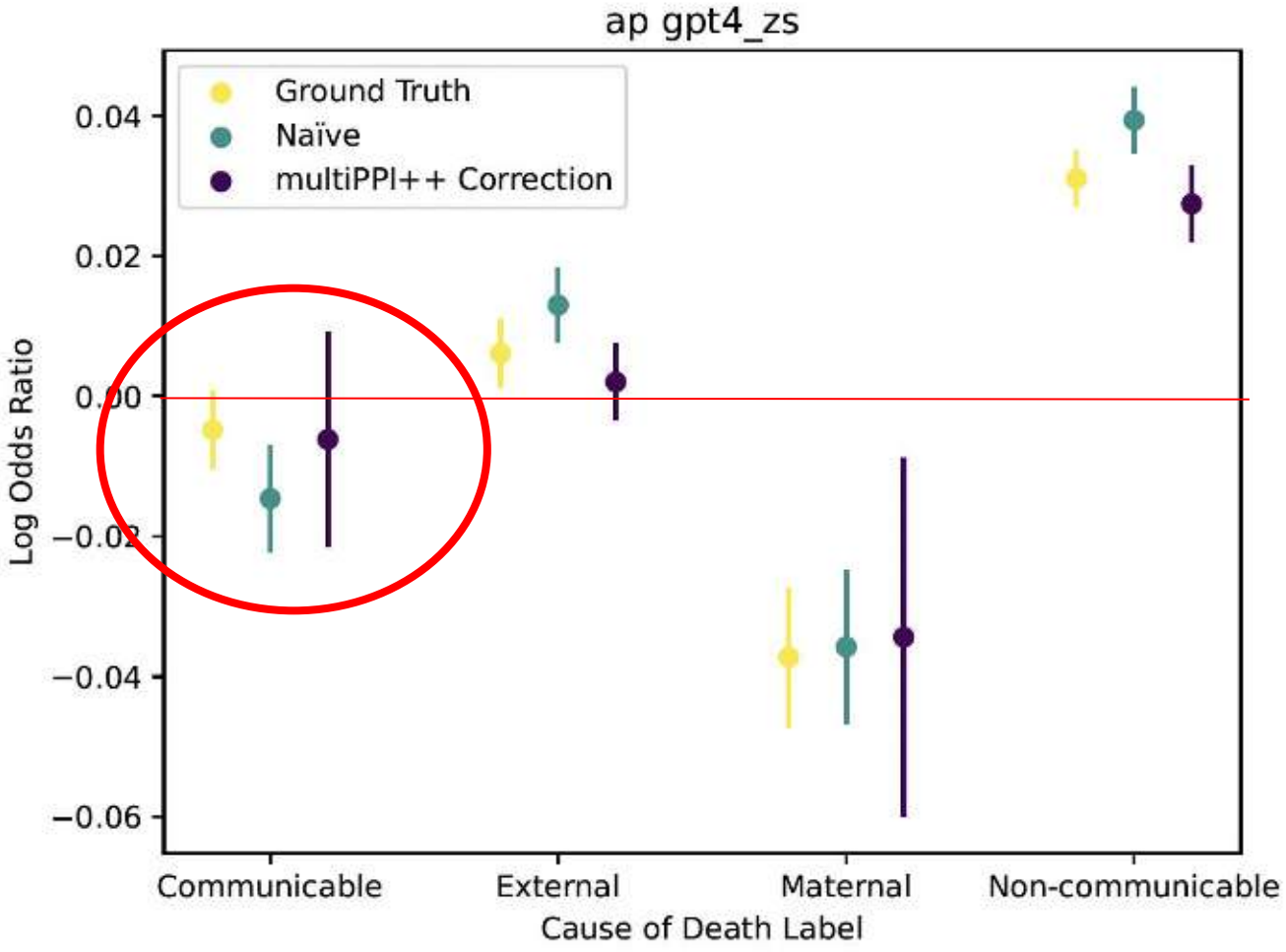


Correcting inference for predictions

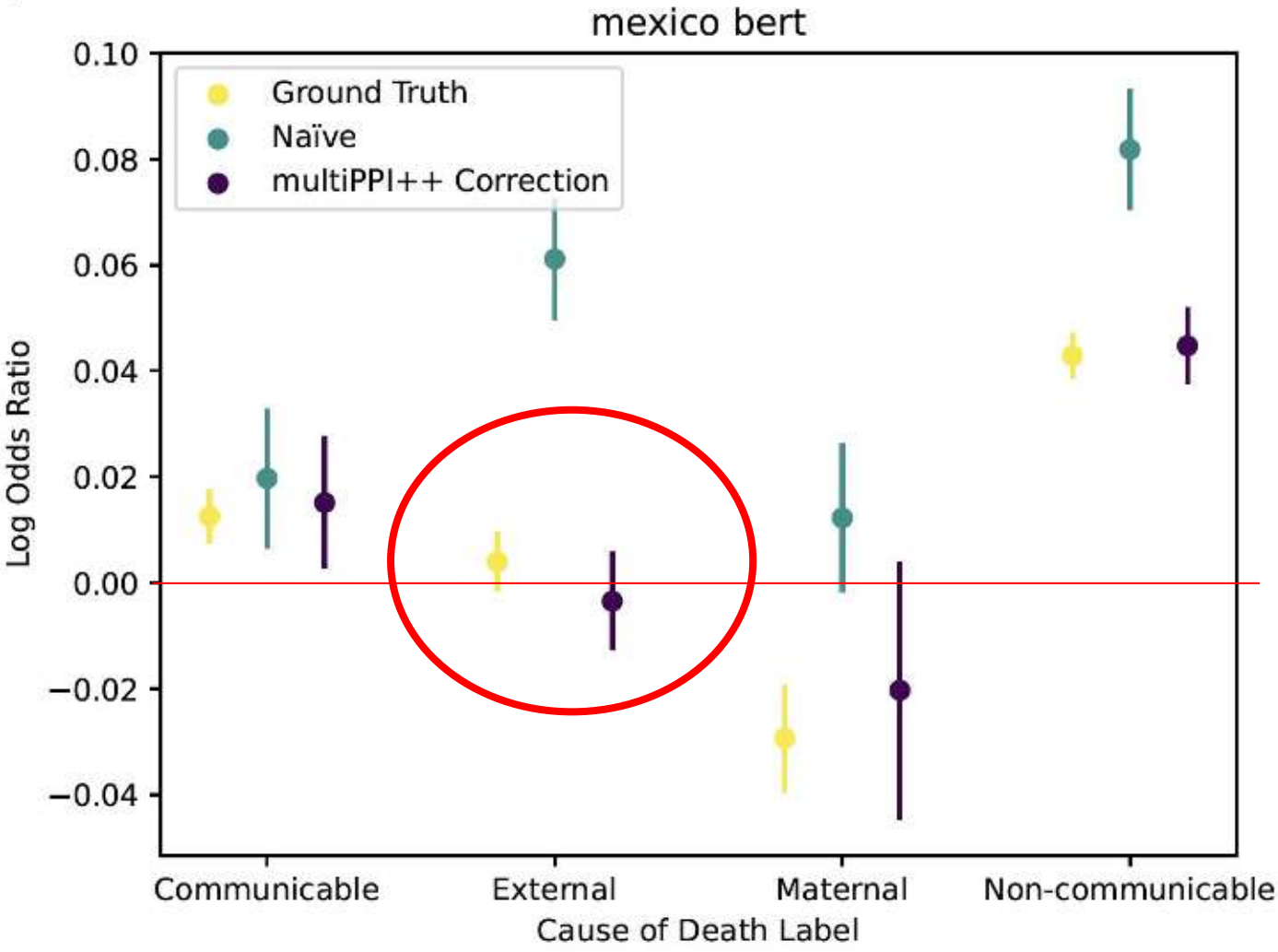


Correcting inference for predictions

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

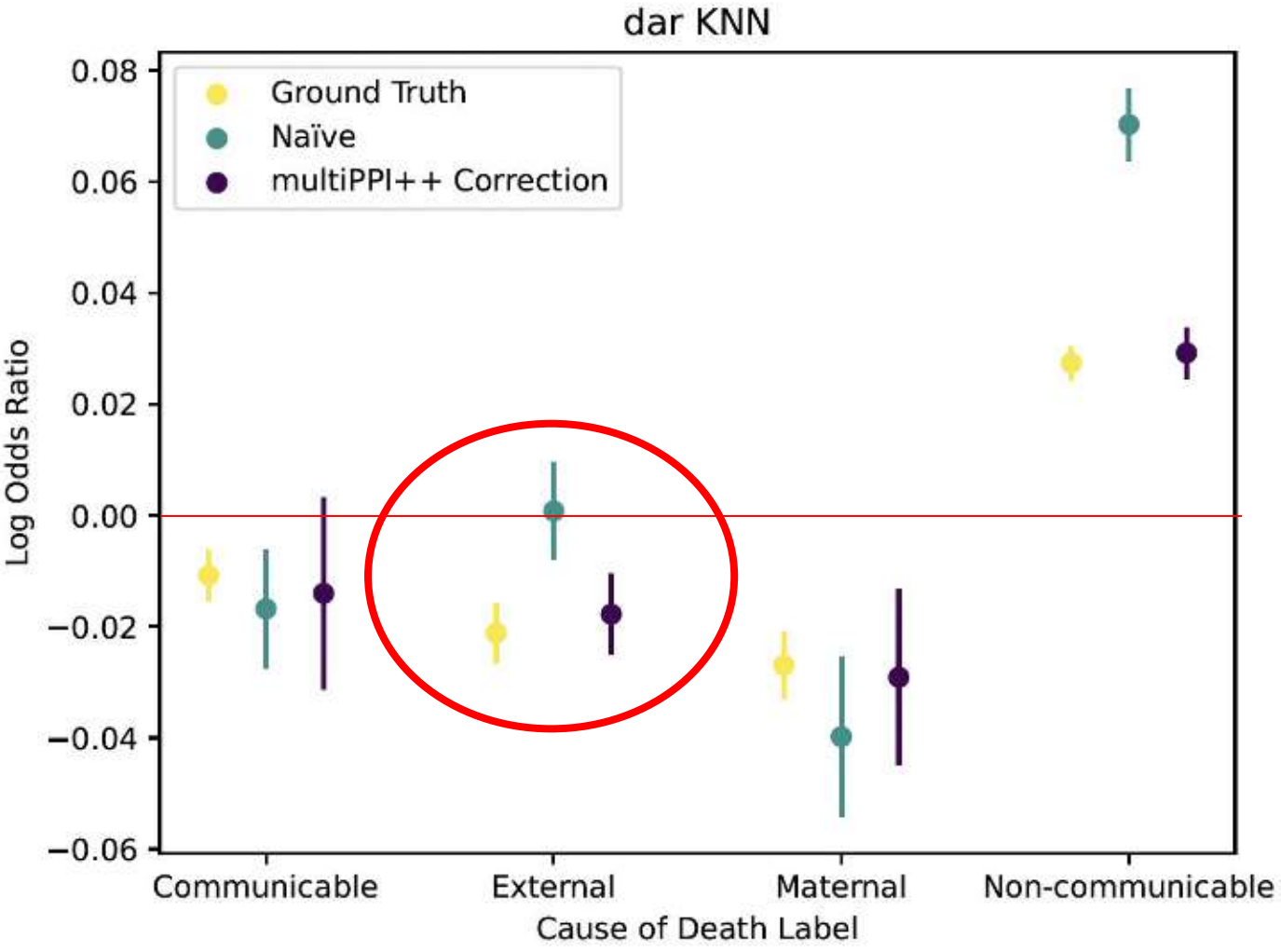


Predictions using generative AI (gpt)



Conclusions change dramatically!!!

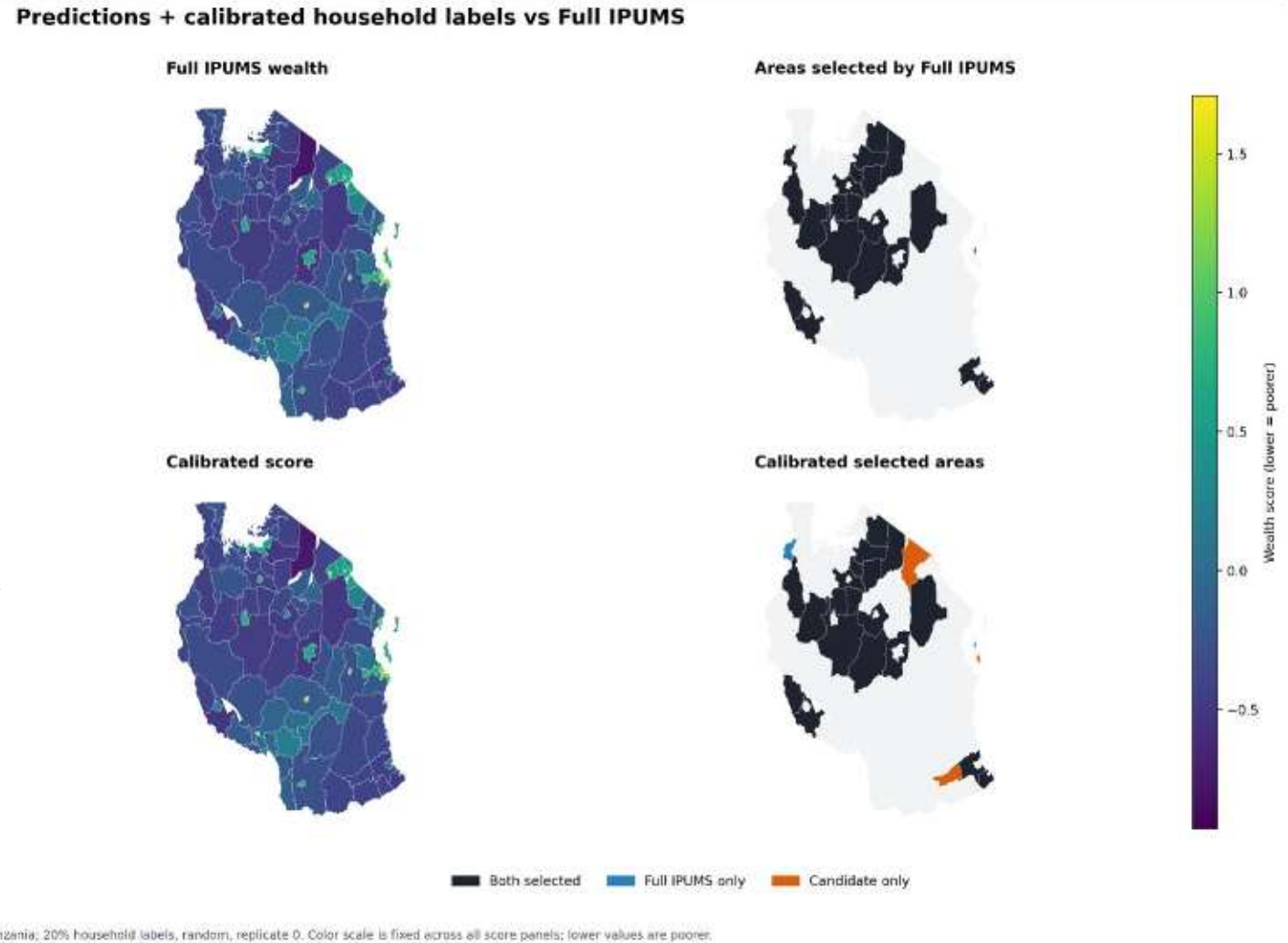
This time with BERT.



We now recover the effect!

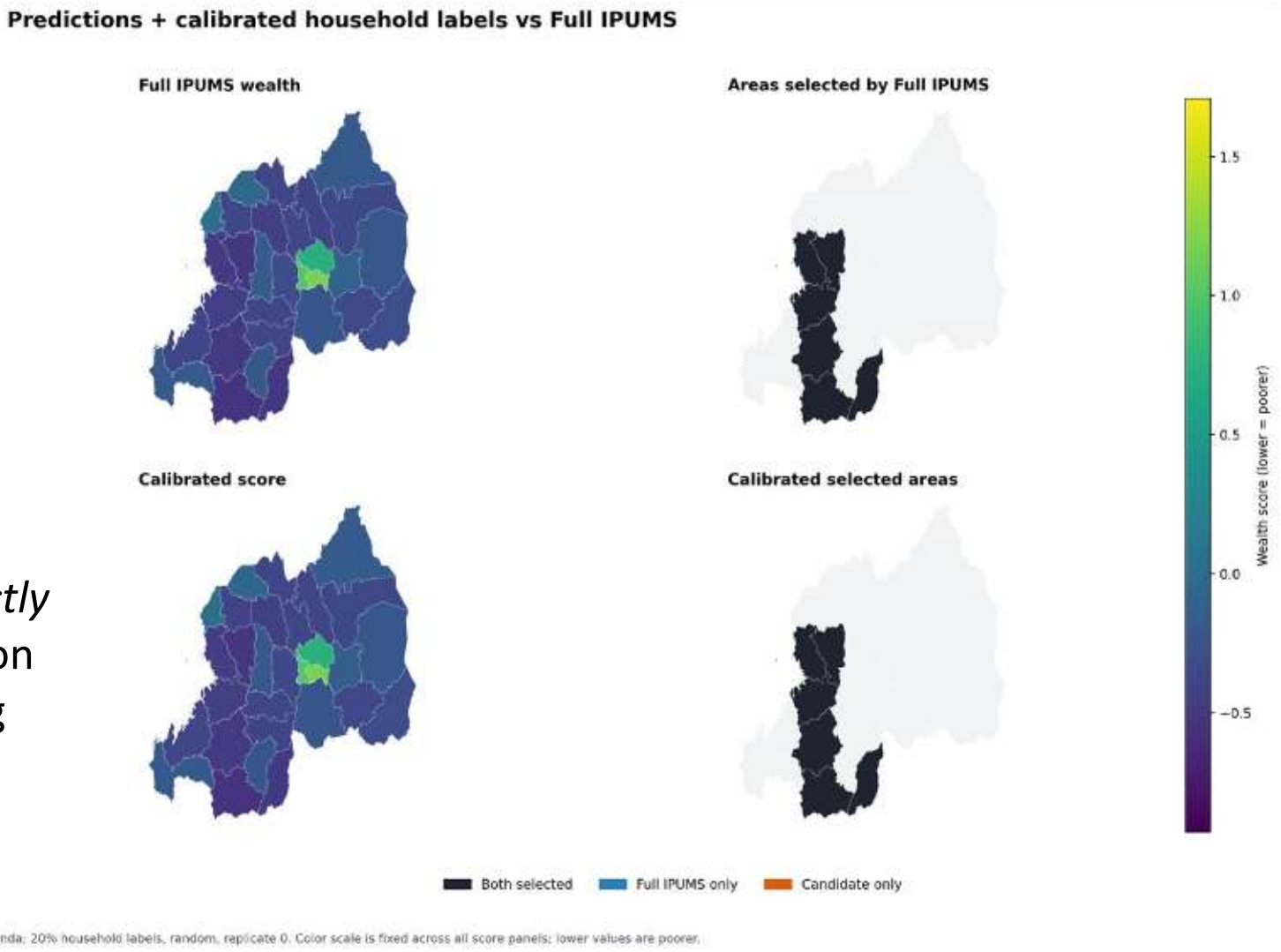
Zambia, w/ calibration

Calibrating bias in prediction model scores with small quantities of local data leads to nearly the same allocation decision as having complete data.



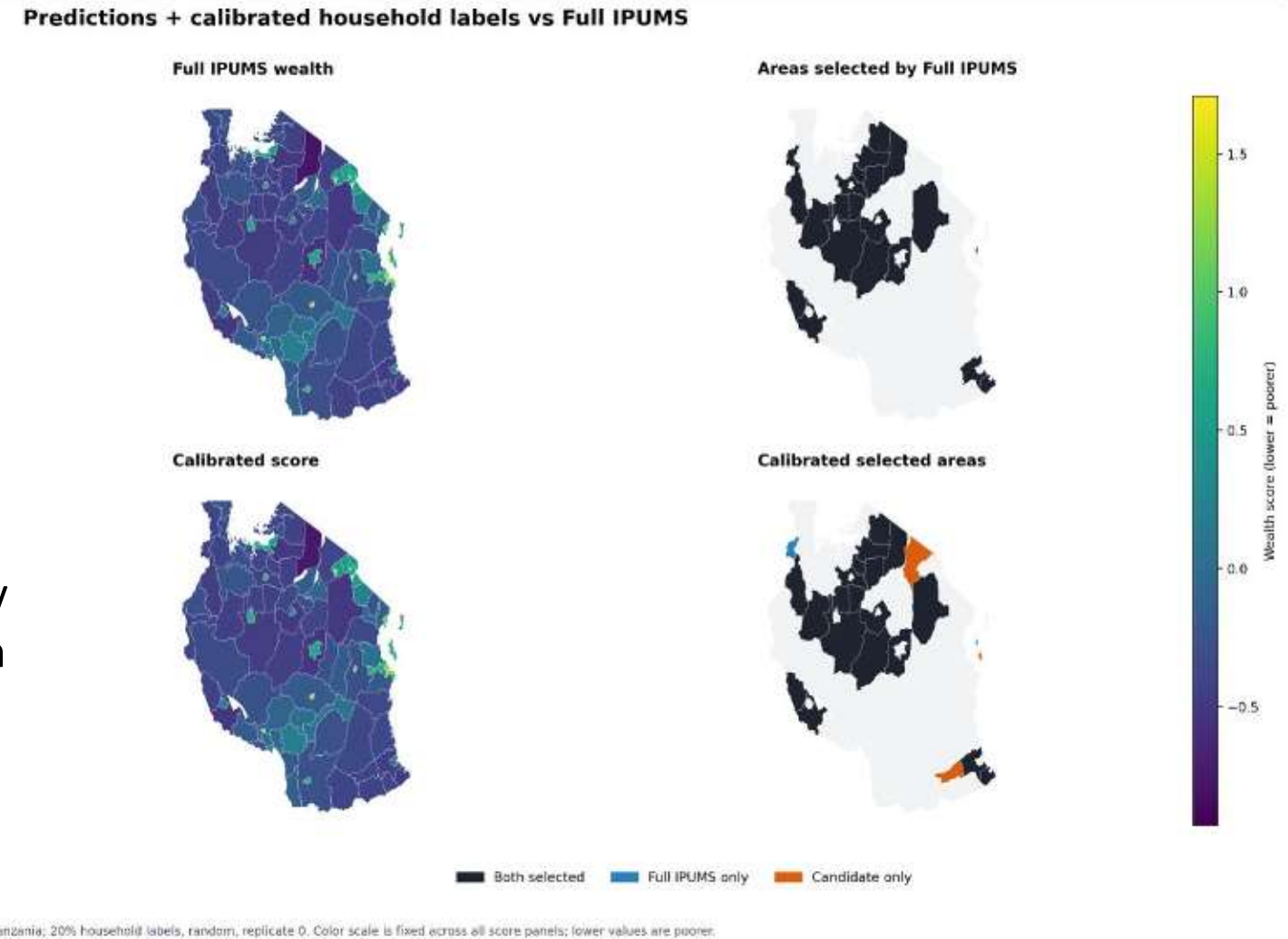
Rwanda, w/ calibration

Calibrating bias in prediction model scores with small quantities of local data leads to *exactly* the same allocation decision as having complete data.



Tanzania, w/ calibration

Calibrating bias in prediction model scores with small quantities of local data leads to nearly the same allocation decision as having complete data.



f(DataYouCanGet)=(DataYouWant)-ish

A little bit of the *right* data can be extremely valuable

Paradigm shift: strategic data collection to maximize learning

From: collect all variables on all observations

To: strategically collect data to augment predictions