Fred Hutch Cancer Center

Inference with Predicted Data

What do we do after we have machine learned (or AI'd?) everything?

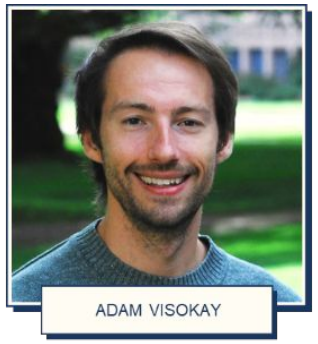
Stephen Salerno
Postdoctoral Researcher
Biostatistics Program, Public Health Sciences

October 15, 2025

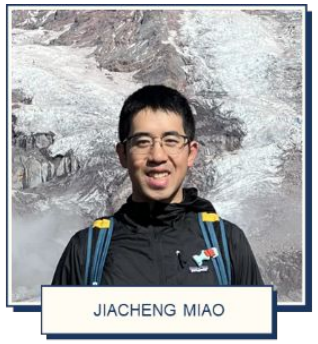
First, Some Acknowledgements


KENTARO HOFFMAN


AWAN AFIAZ


ADAM VISOKAY


SHUXIAN FAN


JIACHENG MIAO


JIANHUI GAO


ANNA NEUFELD

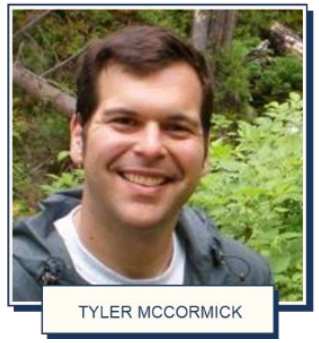

LI LIU


SASHA JOHFRE


DAVID CHENG


JESSE GRONSBELL


QIONGSHI LU


TYLER MCCORMICK

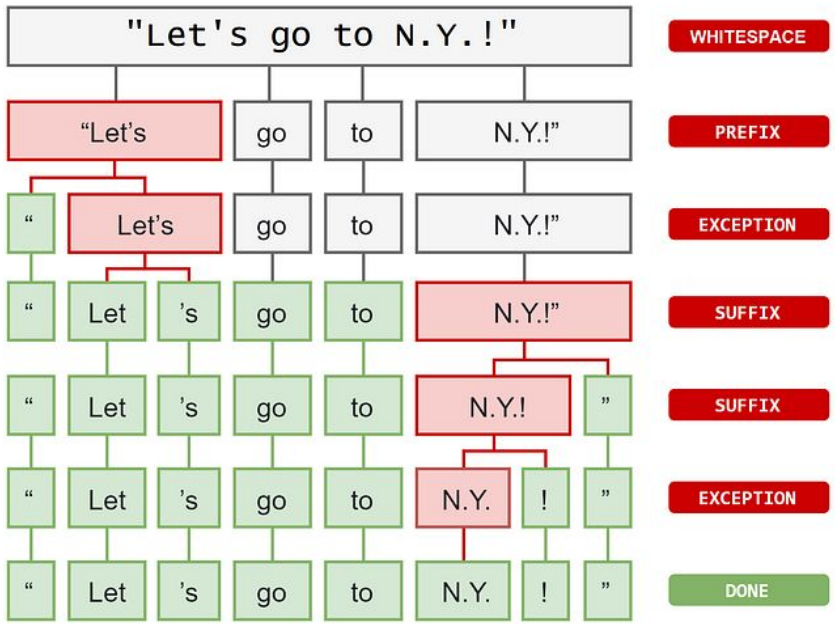

JEFF LEEK

This slide is by no means exhaustive, but it represents individuals who have contributed most recently to the papers I will be highlighting in this talk.

What is AI?

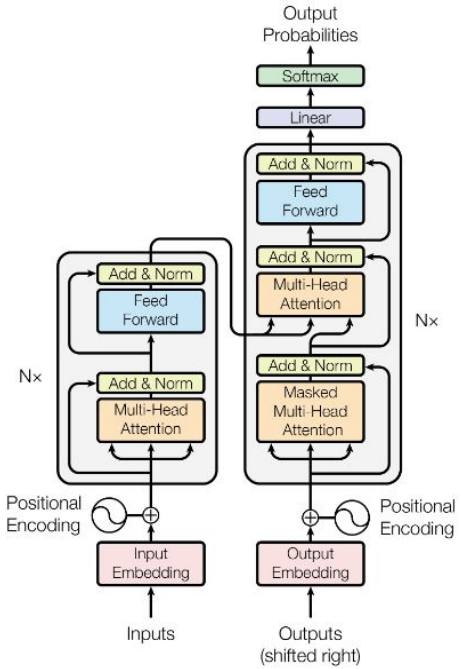
What is AI? A three part definition

Data



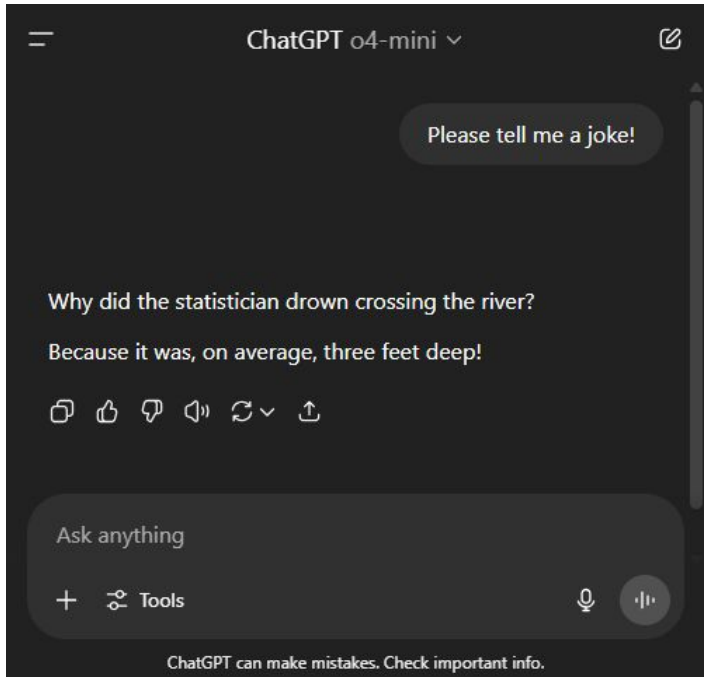
https://www.innerdoc.com/periodic-table-of-nlp-tasks/14-tokenization/

Algorithm



https://arxiv.org/abs/1706.03762

Interface



https://chatgpt.com/

https://simplystatistics.org/posts/2017-01-19-what-is-artificial-intelligence/

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

Why is AI all over the news?

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

How attention changed the game in AI (2017)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

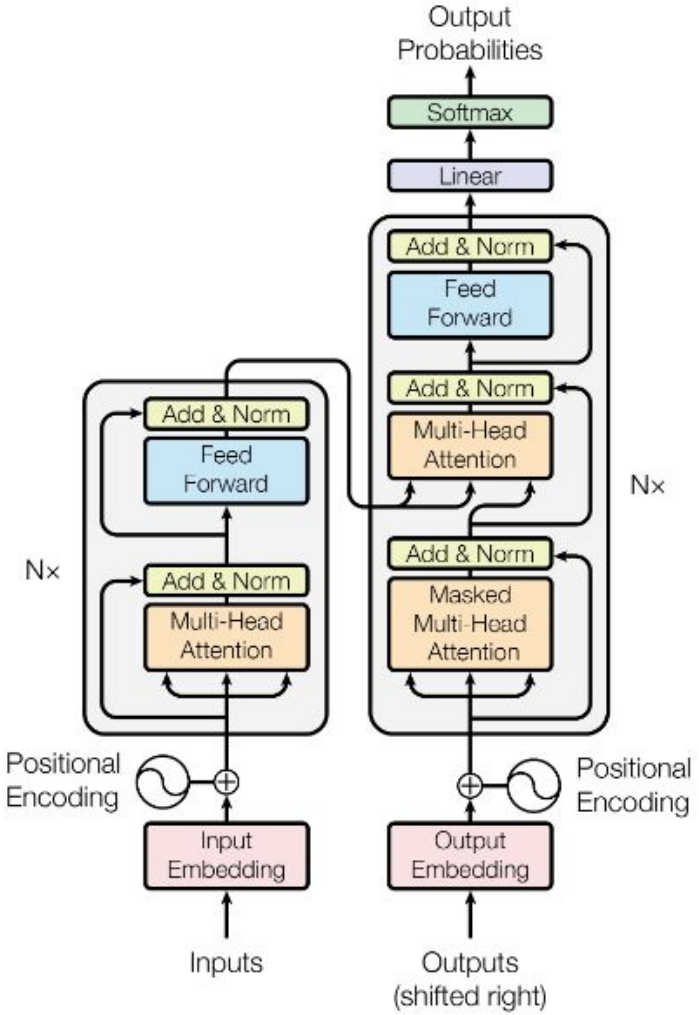
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

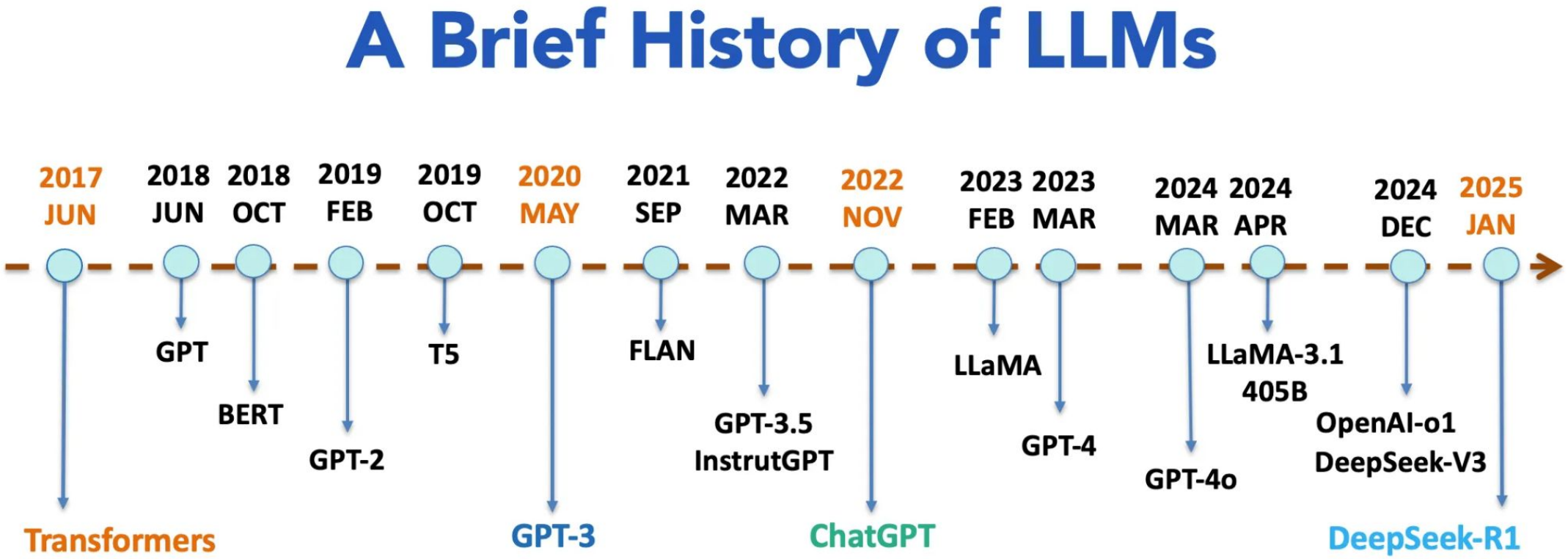
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf



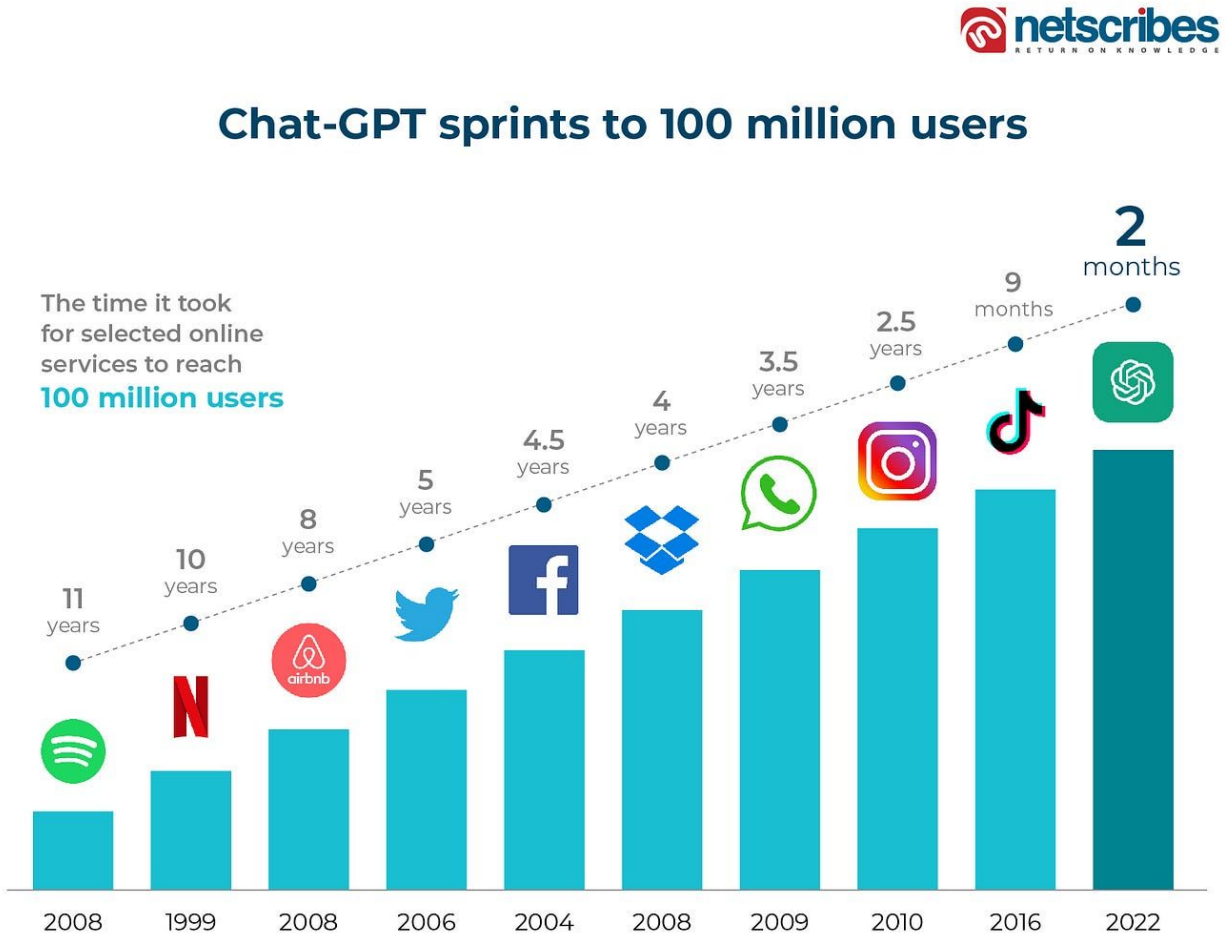
Adapted from *Post-prediction inference*. Source: jtleek.com/talks

LLM breakthroughs have driven AI's current momentum



https://medium.com/@lmpo/a-brief-history-of-lmms-from-transformers-2017-to-deepseek-r1-2025-dae75dd3f59a

AI is everywhere, and it is spreading faster than ever

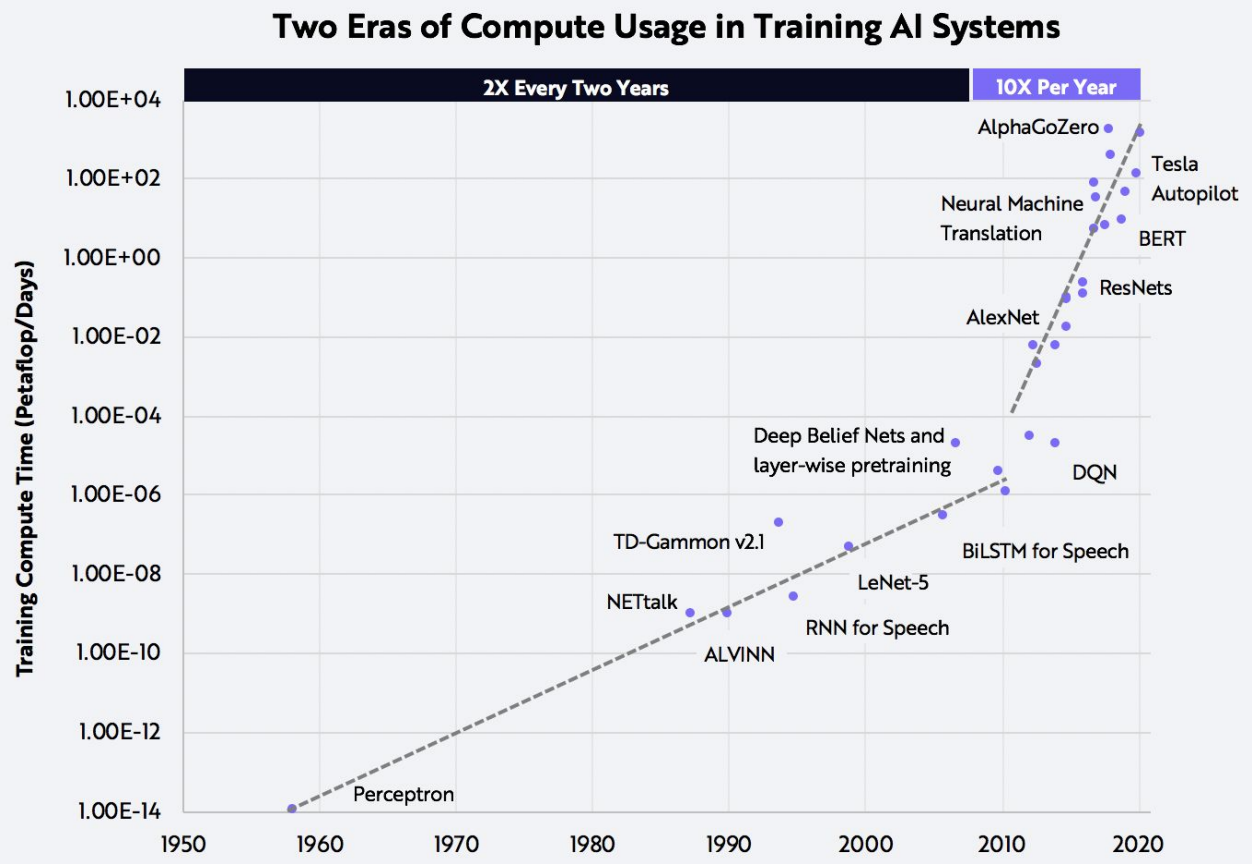


https://www.nebuly.com/blog/2024-the-year-of-llm-user-intelligence

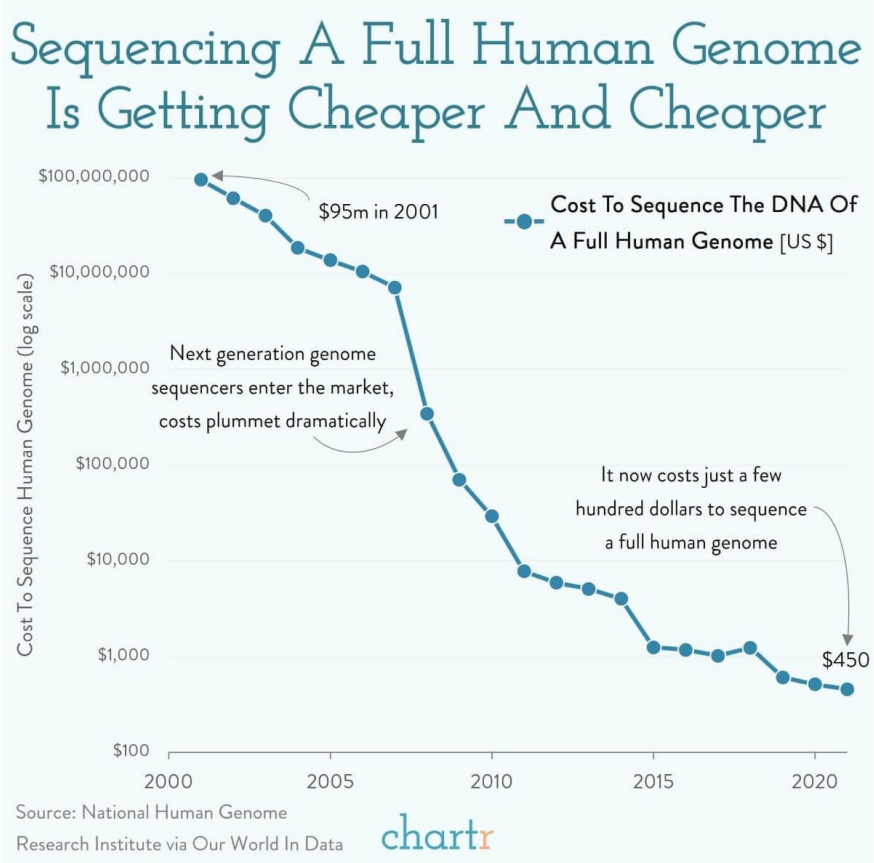
Why is now different, and what does that mean for us?

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

Simultaneous Revolutions in Computing and Biology



https://www.ark-invest.com/articles/analyst-research/ai-training

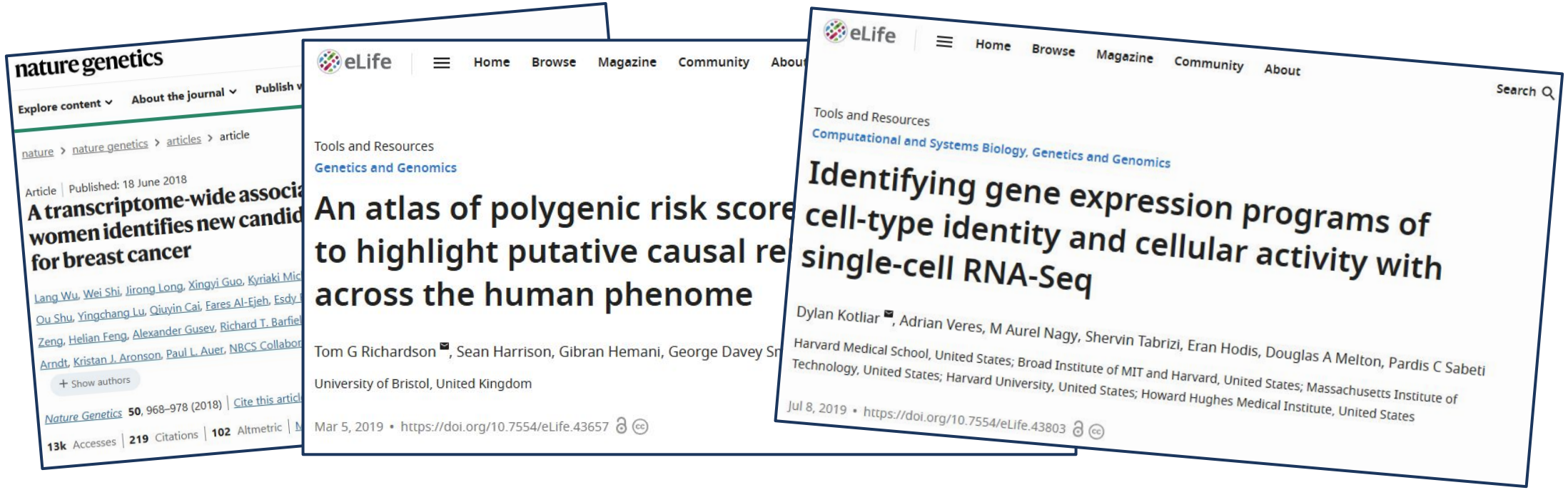


https://digitalfoodlab.com/graph-week-genome-sequencing-price-going/

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

We can now predict complex genetic traits with AI/ML

PRS for disease susceptibility, imputed expression from genotype arrays, TWAS, ...



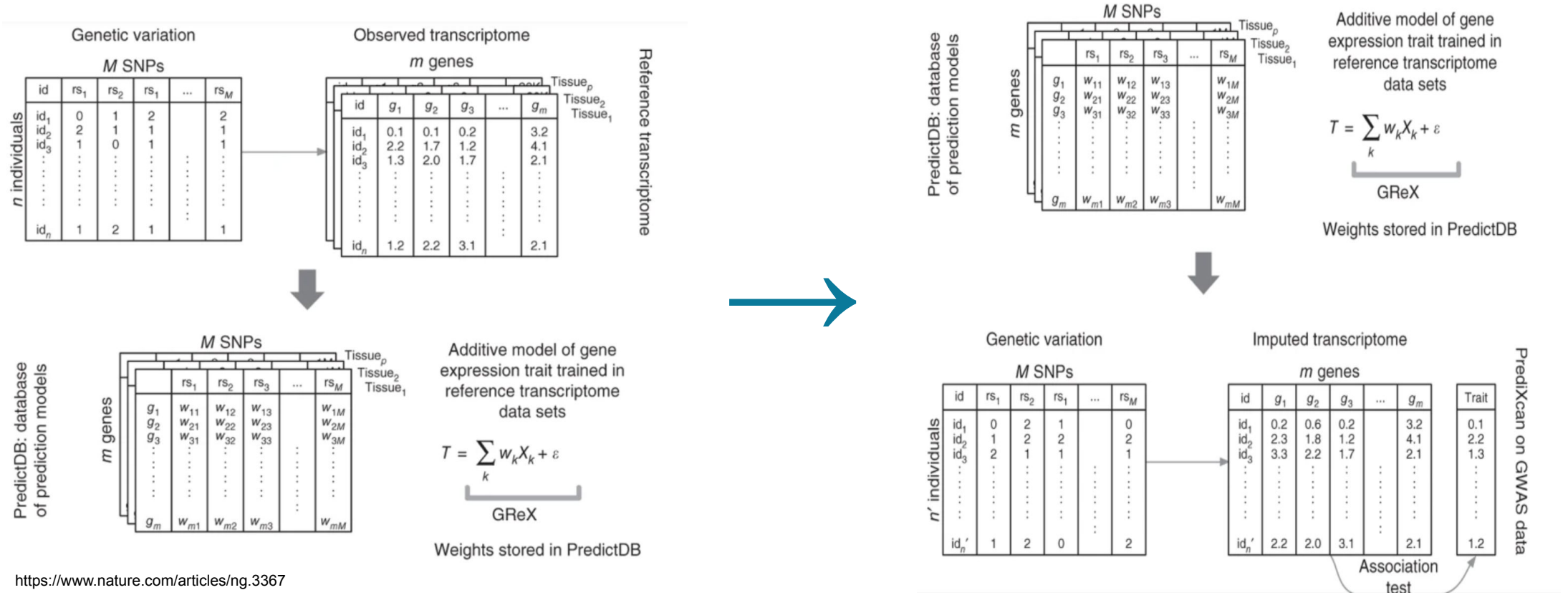
https://www.nature.com/articles/s41588-018-0132-x

https://elifesciences.org/articles/43657

https://elifesciences.org/articles/43803

Example: Common two-stage workflow of a TWAS

1) Model training in a reference panel → 2) Expression imputation and association testing in data



https://www.nature.com/articles/ng.3367

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

How does this relate to statistical inference?

A simple sample size calculation

N = Sample Size

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

How does this relate to statistical inference?

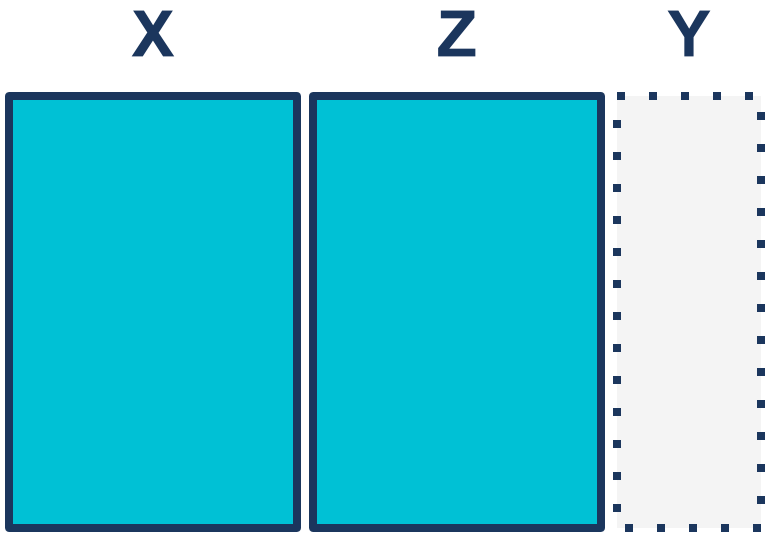
A simple sample size calculation

$$N = \frac{\$\ \text{You Have}}{\$\ \text{Per Sample}}$$

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

In many cases, unlabeled features are cheap and easy to collect

Certain outcomes of interest, however, can be time-consuming or costly to obtain



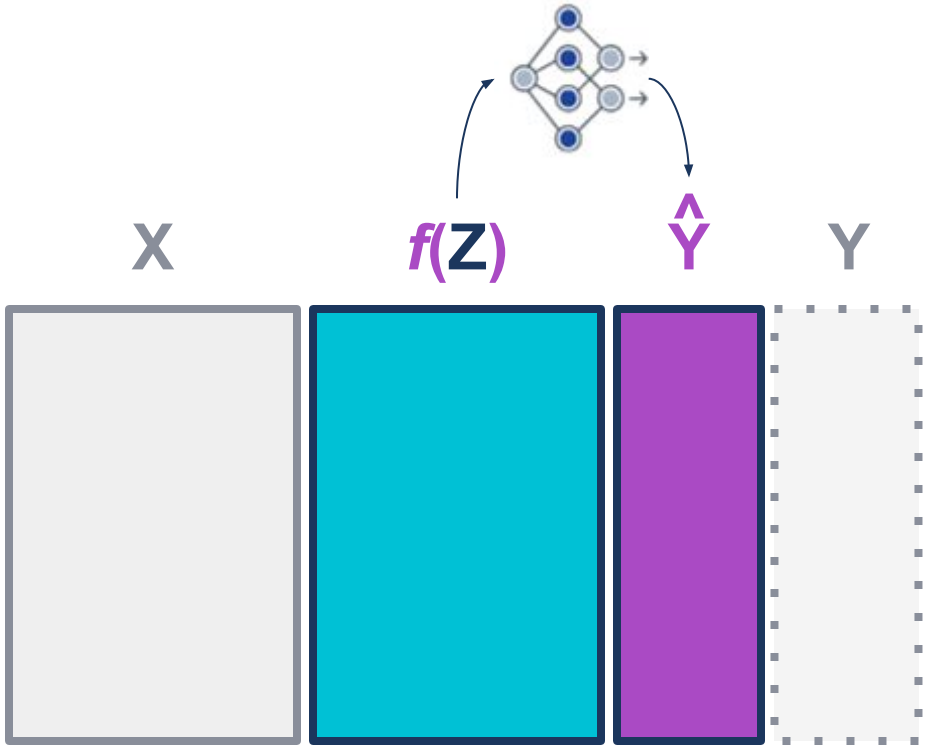
X: Covariates of Interest

Z: Predictive Features

Y: (Missing) Outcome of Interest

f(Data you can get) = (Data you want)-*ish*

Modern studies use accessible data and AI/ML to predict hard-to-measure outcomes



X: Covariates of Interest

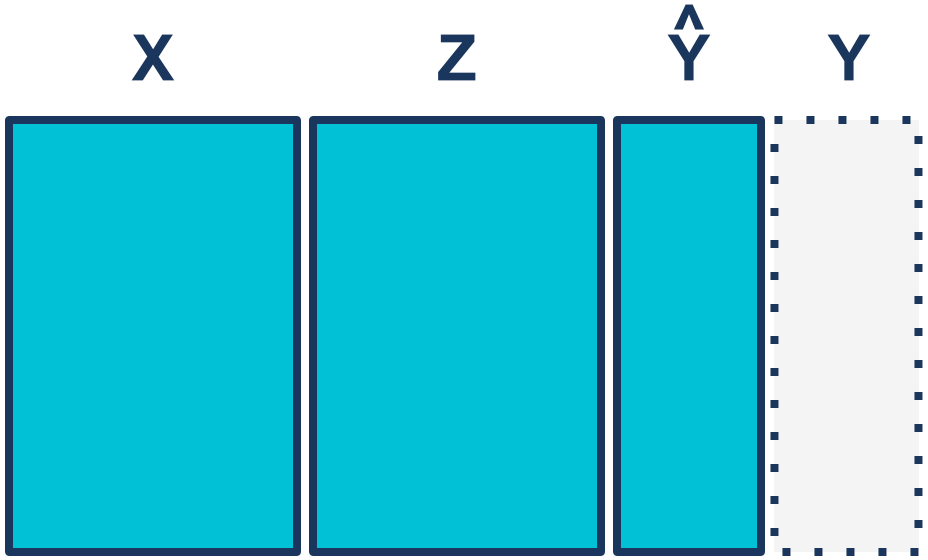
Z: Predictive Features

Y: (Missing) Outcome of Interest

Ŷ: Predicted Outcome

AI/ML-generated predictions are then treated as measured data

These predicted outcomes are used in downstream analyses or in policy decision-making



X: Covariates of Interest

Z: Predictive Features

Y: (Missing) Outcome of Interest

Ŷ: Predicted Outcome

Using predictions as surrogates is appealing…

… and it is happening everywhere

But when our goal is downstream inference/hypothesis testing

This can cause problems

$$g(E[Y \mid X]) = \square_0 + \square_1 X$$

gene expression

drug dose

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

But when our goal is downstream inference/hypothesis testing

This can cause problems

$$g(E[\hat{Y} \mid X]) = \square_0 + \square_1 X$$

predicted gene expression

drug dose

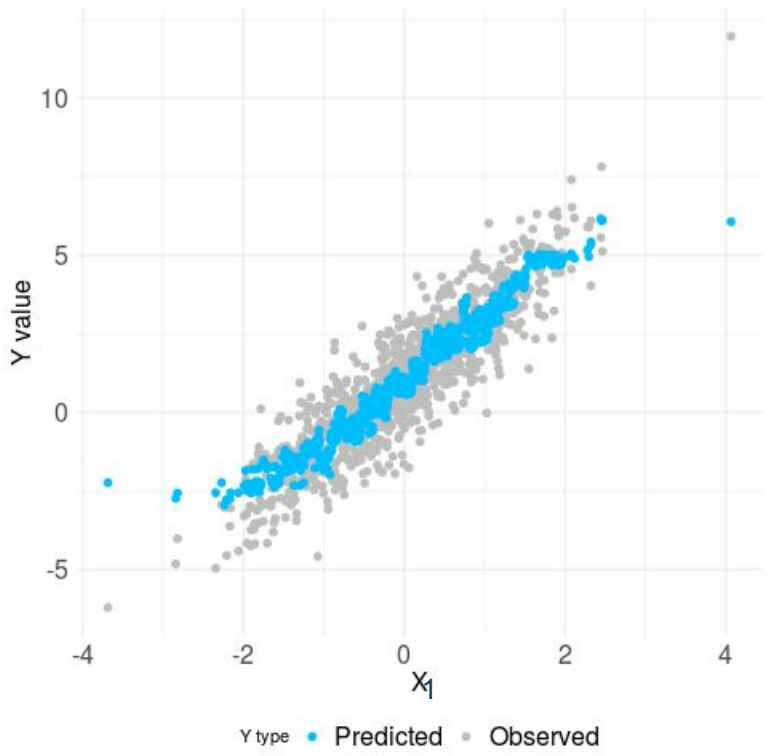
Adapted from *Post-prediction inference*. Source: jtleek.com/talks

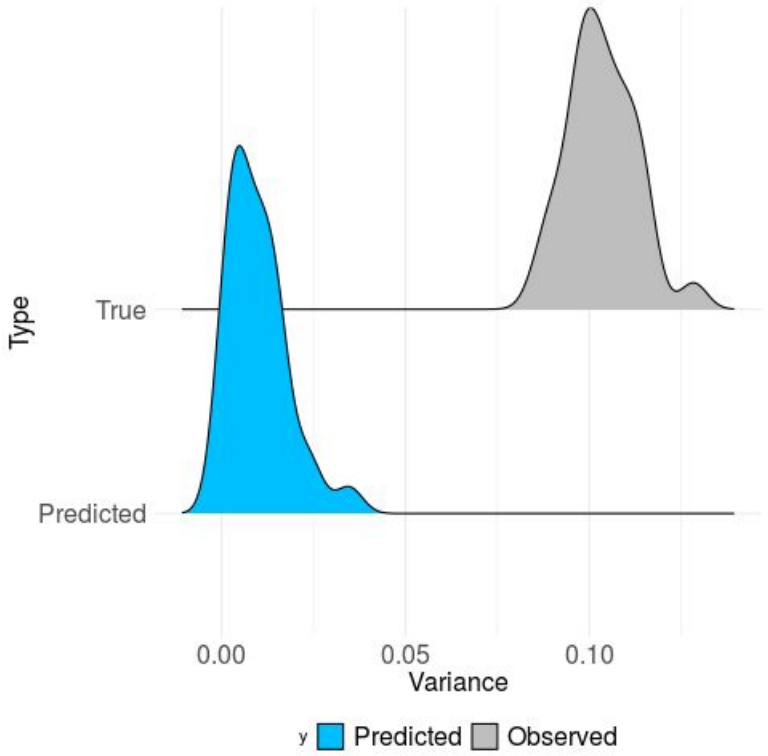
Where can this go wrong?

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

High predictive accuracy ≠ good information for inference

Bias & inability to propagate uncertainty from the predictions to the downstream model





Simulation Setup:

$X_1 \sim N(0,1)$

$X_2 \sim N(0,1)$

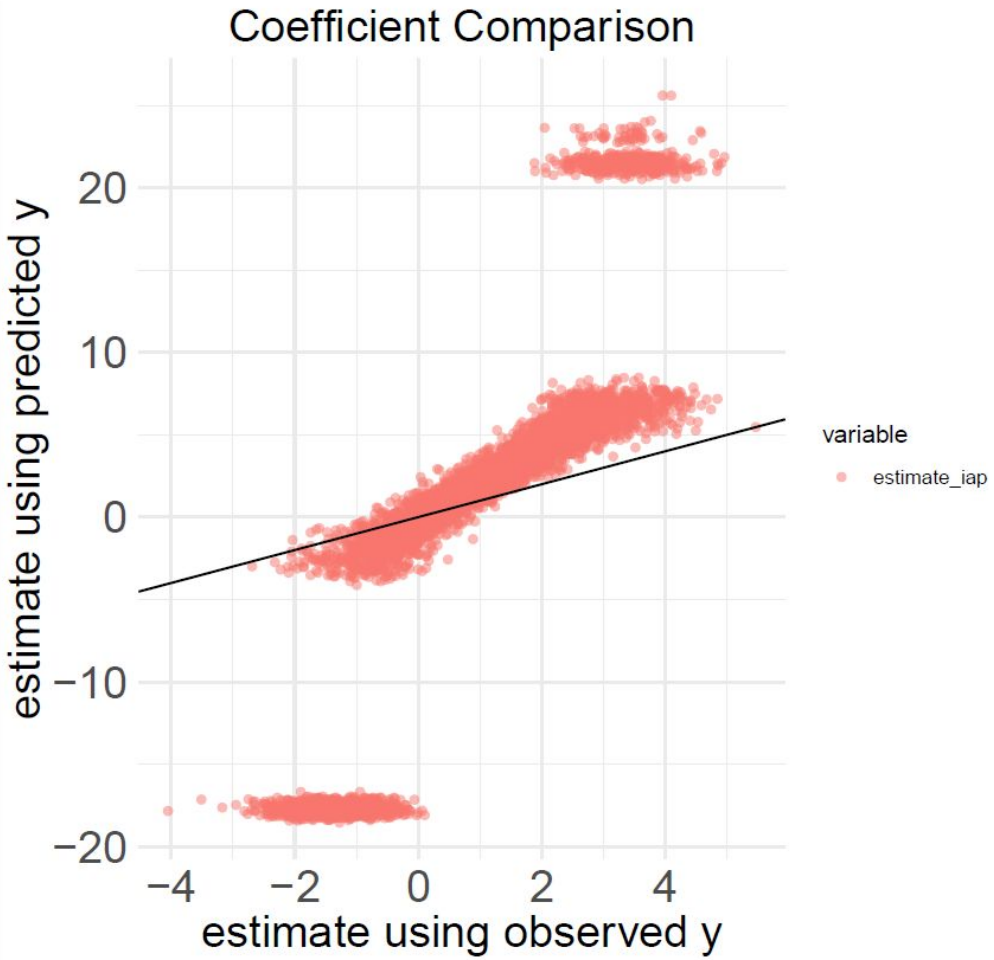
$\mu = 1 + \gamma X_1$

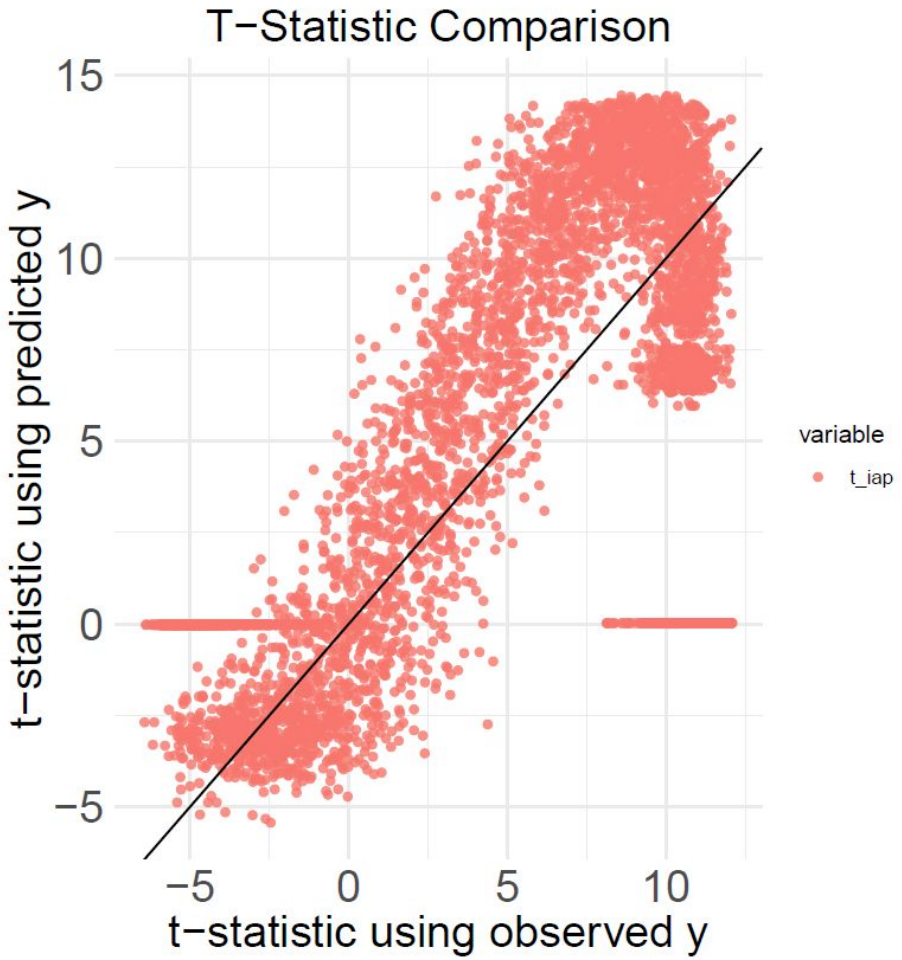
$Y \sim N(\mu, 1)$

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

It gets worse when we compare coefficients/test statistics





Adapted from *Post-prediction inference*. Source: jtleek.com/talks

Where does this leave us?

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

Many AI/ML models are often “black boxes”

We don’t have access to their operating characteristics or training data

Matters arising

Transparency and reproducibility in artificial intelligence

https://doi.org/10.1038/s41586-020-2766-y
Received: 1 February 2020
Accepted: 10 August 2020
Check for updates

Benjamin Haibe-Kains[1,2,3,4,5], George A
Farnoosh Khodakarami[1,2], Massive Anal
Directors*, Levi Waldron[8], Bo Wang[2,3,5,
Anshul Kundaje[13,14], Casey S. Greene[15,16
Jeffrey T. Leek[18], Keegan Korthauer[19,20],
Robert Tibshirani[25,26], Trevor Hastie[25,26],
& Hugo J. W. L. Aerts[6,7,33,34]

ARISING FROM S. M. McKinney et al. *Natur*

https://www.nature.com/articles/s41586-020-2766-y

Table 1 | Essential hyperparameters for reproducing the study for each of the three models

	Lesion	Breast	Case
Learning rate	Missing	0.0001	Missing
Learning rate schedule	Missing	Stated	Missing
Optimizer	Stochastic gradient descent with momentum	Adam	Missing
Momentum	Missing	Not applicable	Not applicable
Batch size	4	Unclear	2
Epochs	Missing	120,000	Missing

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

This is hard to model correctly

And it depends on our upstream (black box) prediction algorithm

$$g(E[\hat{Y} \mid X]) = \square_0 + \square_1 X$$

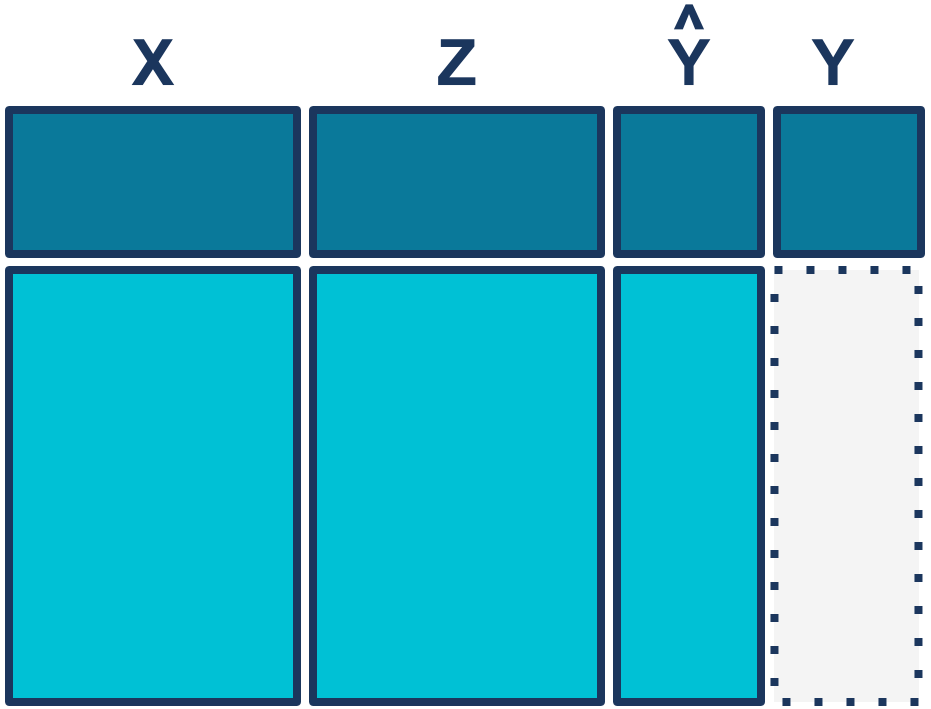
predicted gene expression

drug dose

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

We need methods for “inference with predicted data” (IPD)

Leverage a subset of data with gold-standard outcomes to calibrate inference in the larger study



: Labeled : Unlabeled

X: Covariates of Interest

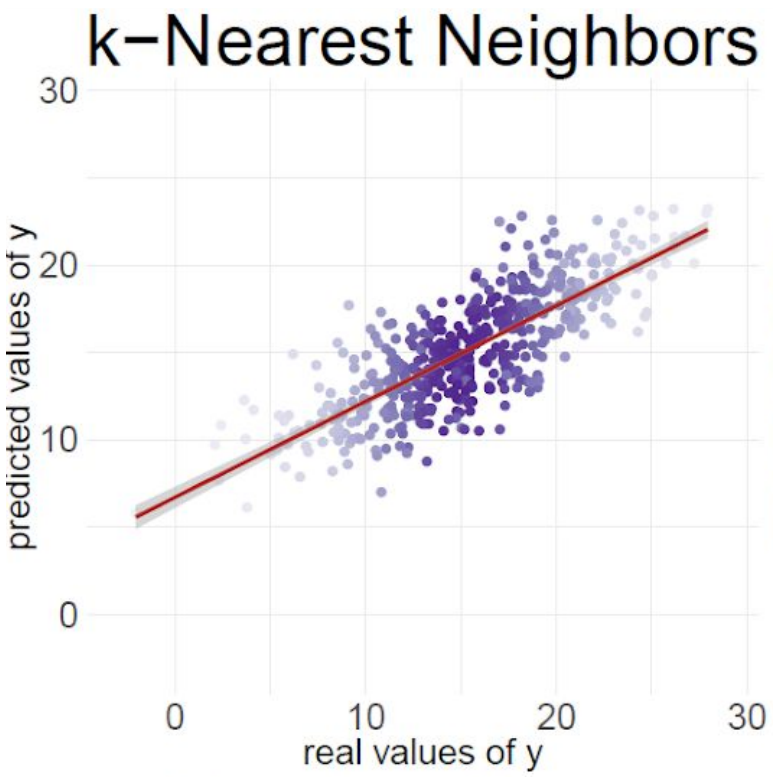
Z: Predictive Features

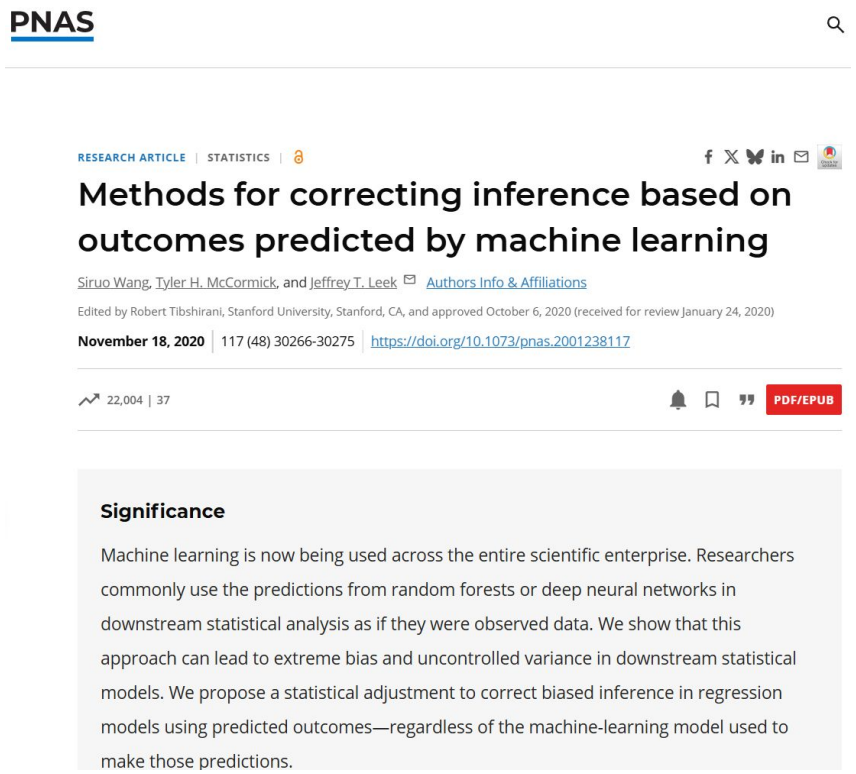
Y: (Missing) Outcome of Interest

Ŷ: Predicted Outcome

Post-Prediction Inference (Wang et al., 2020)

A key observation that the predicted and true outcomes often have a simple relationship





PNAS

RESEARCH ARTICLE | STATISTICS

Methods for correcting inference based on outcomes predicted by machine learning

Siruo Wang, Tyler H. McCormick, and Jeffrey T. Leek Authors Info & Affiliations

Edited by Robert Tibshirani, Stanford University, Stanford, CA, and approved October 6, 2020 (received for review January 24, 2020)

November 18, 2020 | 117 (48) 30266-30275 | https://doi.org/10.1073/pnas.2001238117

22,004 | 37 PDF/EPUB

Significance

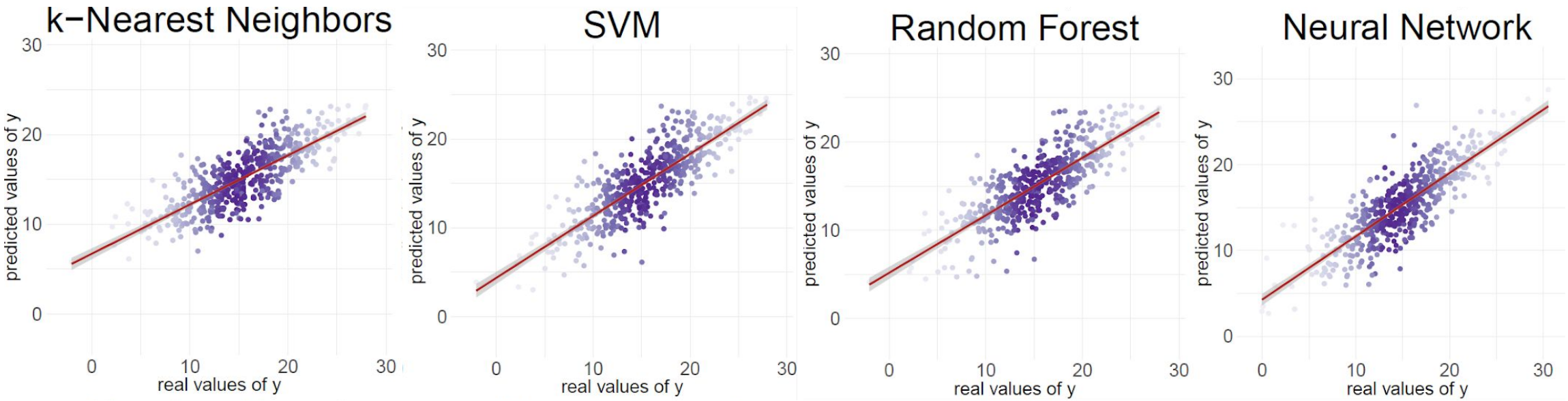
Machine learning is now being used across the entire scientific enterprise. Researchers commonly use the predictions from random forests or deep neural networks in downstream statistical analysis as if they were observed data. We show that this approach can lead to extreme bias and uncontrolled variance in downstream statistical models. We propose a statistical adjustment to correct biased inference in regression models using predicted outcomes—regardless of the machine-learning model used to make those predictions.

https://www.pnas.org/doi/abs/10.1073/pnas.2001238117

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

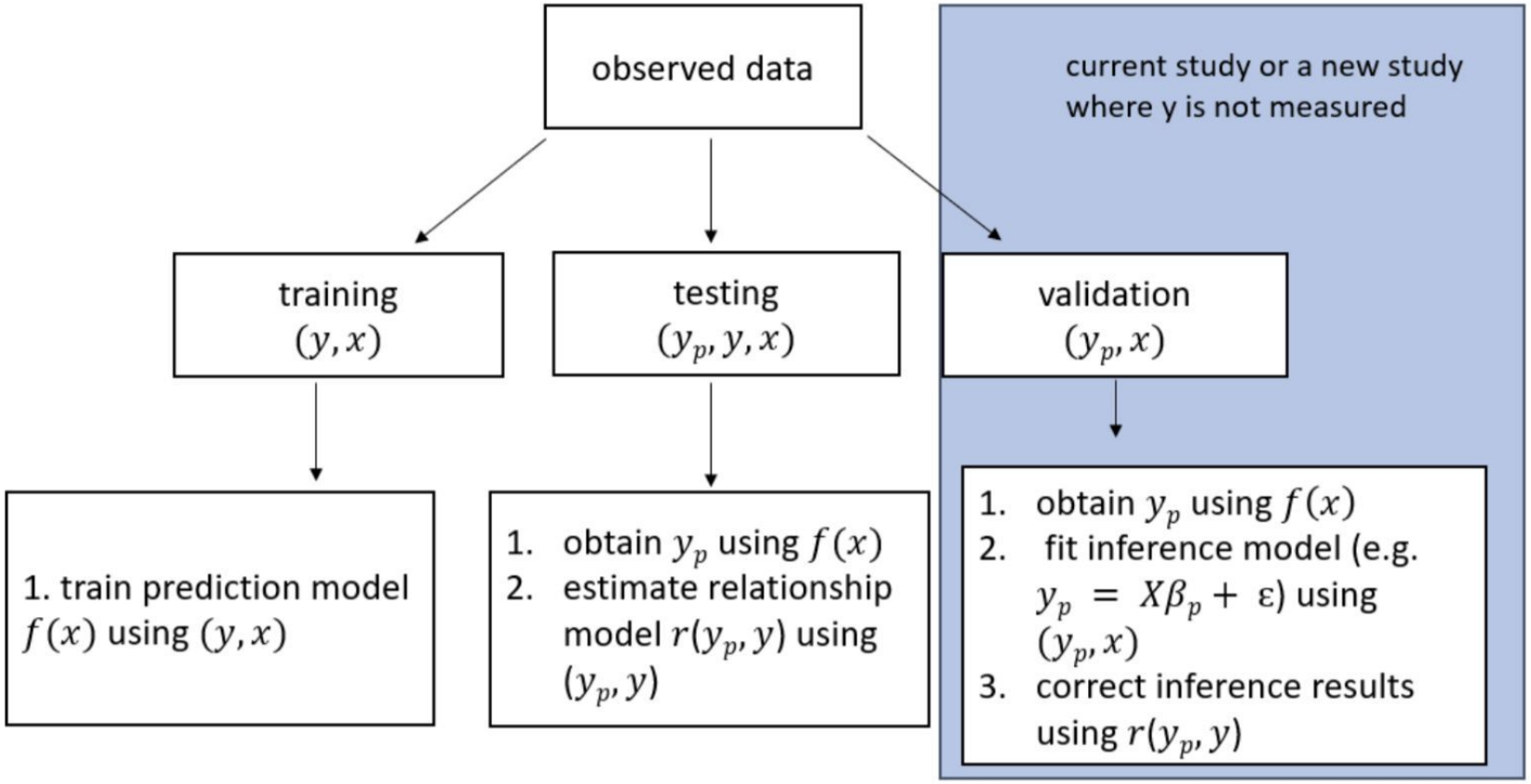
Predicted and true outcomes often have a simple relationship

Regardless of the predictive model used



Adapted from *Post-prediction inference*. Source: jtleek.com/talks

Post-Prediction Inference models this relationship in the labeled data



y: observed outcome x: observed covariate y_p: predicted outcome
$f(x)$: prediction model $r(y_p, y)$: relationship model

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

PostPI uses the testing set to correct inference in the validation set

Consider a downstream linear regression

$$Y = \boldsymbol{X}^\top \beta + \varepsilon, \quad \mathbb{E}[\varepsilon \mid \boldsymbol{X}] = 0.$$

Fit a linear relationship model in the labeled set:

$$Y_{\mathcal{L}} = \gamma_0 + \gamma_1 f(\boldsymbol{Z}_{\mathcal{L}}) + \eta, \quad \mathbb{E}[\eta \mid f(\boldsymbol{Z}_{\mathcal{L}})] = 0.$$

then form pseudo-outcomes in the unlabeled set:

$$Y_{\mathcal{U}}^* = \hat{\gamma}_0 + \hat{\gamma}_1 f(\boldsymbol{Z}_{\mathcal{U}})$$

Their final estimator is:

$$\hat{\beta}_{\text{PostPI}} = \left(\boldsymbol{X}_{\mathcal{U}}^\top \boldsymbol{X}_{\mathcal{U}}\right)^{-1} \boldsymbol{X}_{\mathcal{U}}^\top Y_{\mathcal{U}}^* \qquad (3)$$

For inference, PostPI approximates the conditional variance

$$\mathrm{Var}\left(Y_{\mathcal{U}} \mid \boldsymbol{X}_{\mathcal{U}}\right) \approx \sigma_r^2 + \gamma_1^2 \sigma_p^2,$$

their standard error estimator is

$$SE\left(\hat{\beta} \mid \boldsymbol{X}_{\mathcal{U}}\right) = \sqrt{\left(\boldsymbol{X}_{\mathcal{U}}^{\top} \boldsymbol{X}_{\mathcal{U}}\right)^{-1} \left(\hat{\sigma}_r^2 + \hat{\gamma}_1^2 \hat{\sigma}_p^2\right)}. \tag{5}$$

This mirrors linear regression, with a scalar residual variance and inverse Gram matrix

We propose a moment-based generalization to PostPI

This relaxes several assumptions from the original method

PostPI assumes the prediction error, η, is uncorrelated with the covariates, X. In many realistic settings, η will share structure with X, so the bias is nonzero.

$$\beta = \left[\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)\right]^{-1}\{\gamma_1\mathbb{E}[\boldsymbol{X}f(\boldsymbol{Z})] + \mathbb{E}(\boldsymbol{X}\eta)\}$$
$$\neq \gamma_1\left[\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)\right]^{-1}\mathbb{E}[\boldsymbol{X}f(\boldsymbol{Z})] = \beta_{\text{PostPI}},$$

PostPI's residual variance has two components, one from the relationship model and one from the 'naive' model

$$SE\left(\hat{\beta} \mid \boldsymbol{X}_{\mathcal{U}}\right) = \sqrt{\left(\boldsymbol{X}_{\mathcal{U}}^\top\boldsymbol{X}_{\mathcal{U}}\right)^{-1}\left(\hat{\sigma}_r^2 + \hat{\gamma}_1^2\hat{\sigma}_p^2\right)}.$$

As the precision matrix scales as O(1/N), this term vanishes as the size of the unlabeled set (N) goes to infinity

Our proposal reduces to PostPI if its assumptions hold

but extends to more general settings if it does not, with more theoretical guarantees

Assuming independent $\mathcal{L}$ and $\mathcal{U}$, we estimate $\hat{\gamma}_0$, $\hat{\gamma}_1$, and $\hat{\eta}_i = Y_i - \hat{\gamma}_0 - \hat{\gamma}_1 f(\boldsymbol{Z}_i)$, from the *labeled* set, and covariances from both the *labeled* ($\mathcal{L}$) and *unlabeled* ($\mathcal{U}$) sets, where we define the following empirical moments:

$$\hat{C}^{\mathcal{U}}_{\boldsymbol{X}f} = \frac{1}{N}\sum_{i=n+1}^{n+N} \boldsymbol{X}_i f_i,\ \hat{C}^{\mathcal{L}}_{\boldsymbol{X}\eta} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \hat{\eta}_i,\ \hat{M}^{\mathcal{U}}_{\boldsymbol{X}\boldsymbol{X}} = \frac{1}{N}\sum_{i=n+1}^{n+N} \boldsymbol{X}_i \boldsymbol{X}_i^\top.$$

Then, our extended estimator is given by

$$\hat{\beta} = \left(\hat{M}^{\mathcal{U}}_{\boldsymbol{X}\boldsymbol{X}}\right)^{-1}\left(\hat{\gamma}_1 \hat{C}^{\mathcal{U}}_{\boldsymbol{X}f} + \hat{C}^{\mathcal{L}}_{\boldsymbol{X}\eta}\right). \tag{4}$$

If $\mathbb{E}[\boldsymbol{X}\eta] = 0$, then (4) reduces to the original PostPI estimator. Otherwise, (4) extends PostPI to settings where the prediction error is correlated with $\boldsymbol{X}$.

We also generalize the variance estimation

Consider the following covariance matrices:

$$M = \mathbb{E}(XX^\top), \quad S_1 = \mathrm{Var}\left[Xf(\boldsymbol{Z})\right], \quad S_2 = \mathrm{Var}\left(X\eta\right).$$

We can show that the estimated standard error is

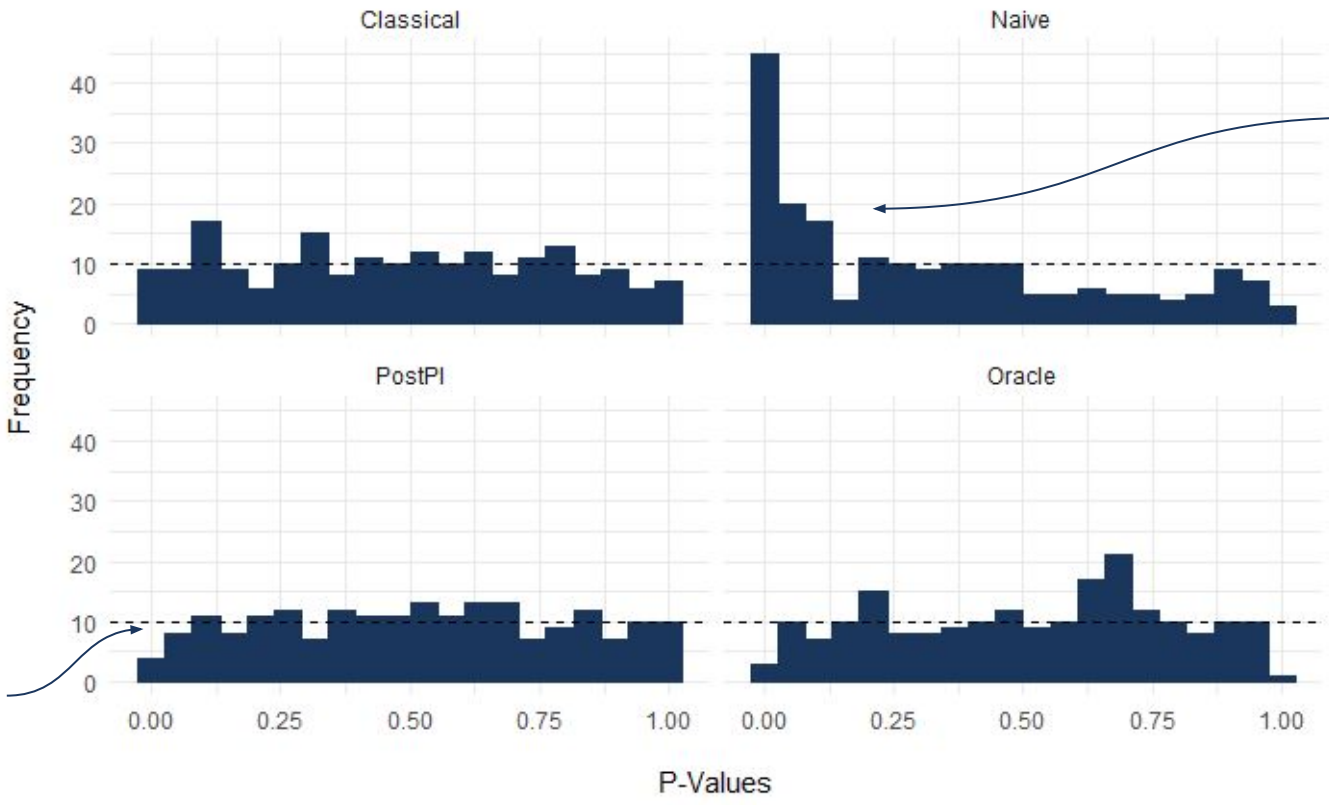
$$\hat{\mathrm{SE}}(\hat{\beta}) = \sqrt{\frac{1}{N}\left(\hat{M}^{\mathcal{U}}_{\boldsymbol{XX}}\right)^{-1}\left(\hat{\gamma}_1^2\hat{S}_1^{\mathcal{U}} + \frac{N}{n}\hat{S}_2^{\mathcal{L}}\right)\left(\hat{M}^{\mathcal{U}}_{\boldsymbol{XX}}\right)^{-1}}, \tag{6}$$

with S_1 and S_2 estimated by their sample analogs in the *unlabeled* and *labeled* sets, respectively. For scalar S_1, S_2, and for $n = N$, (6) reduces to (5). We form Wald-type confidence intervals as $c^\top\hat{\beta} \pm z_{1-\alpha/2}\sqrt{c^\top \mathrm{Var}(\hat{\beta})c}$.

IPD methods calibrate inference for proper Type 1 error control

P-values should be uniform (flat) under the null hypothesis

True Effect = 0



Does not control Type 1 error

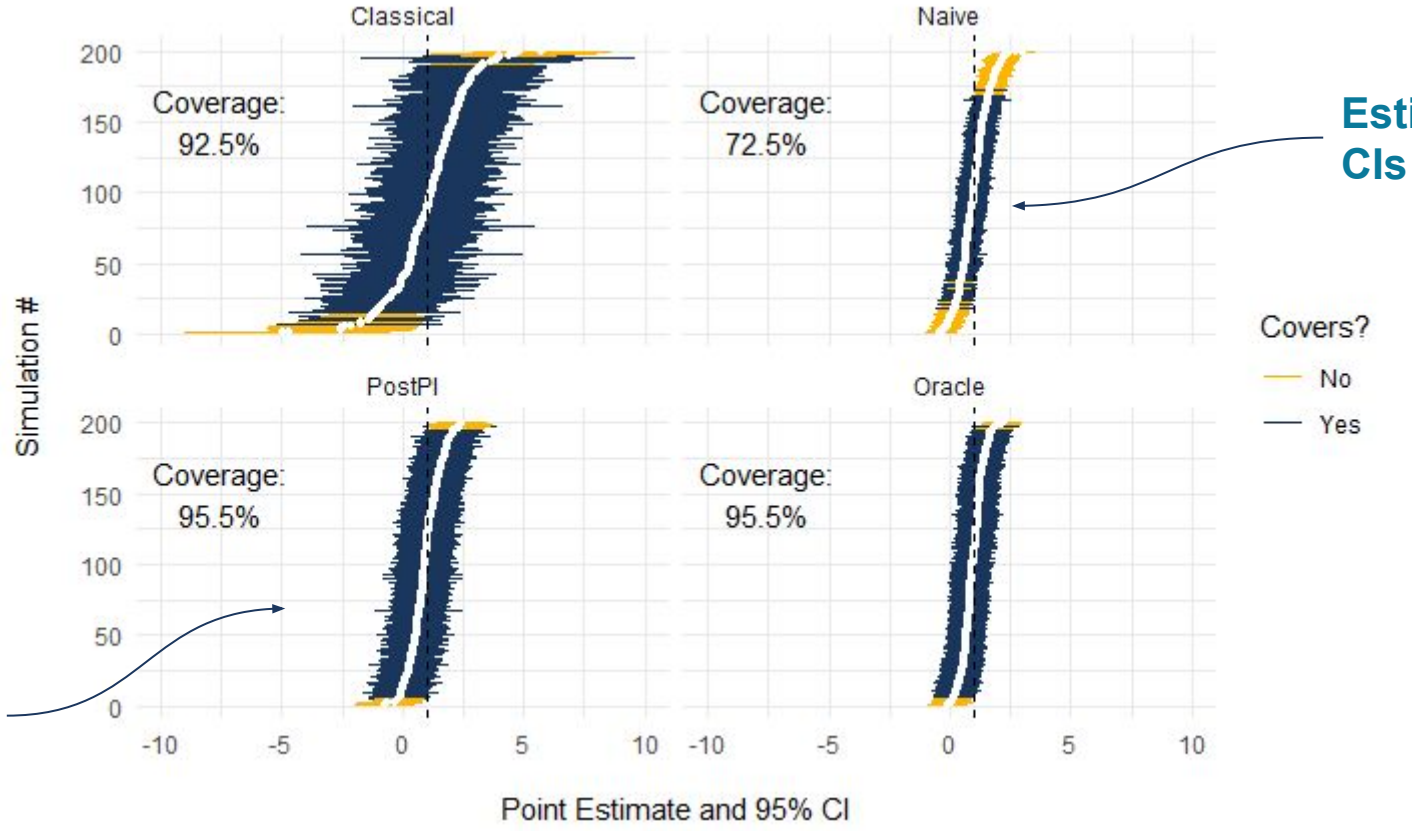
Controls Type 1 error appropriately

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

And nominal coverage (% of CIs that contain the true parameter)

CIs are narrower than complete data analysis, but wider than the ideal, i.e., there is “no free lunch”

True Effect = 1



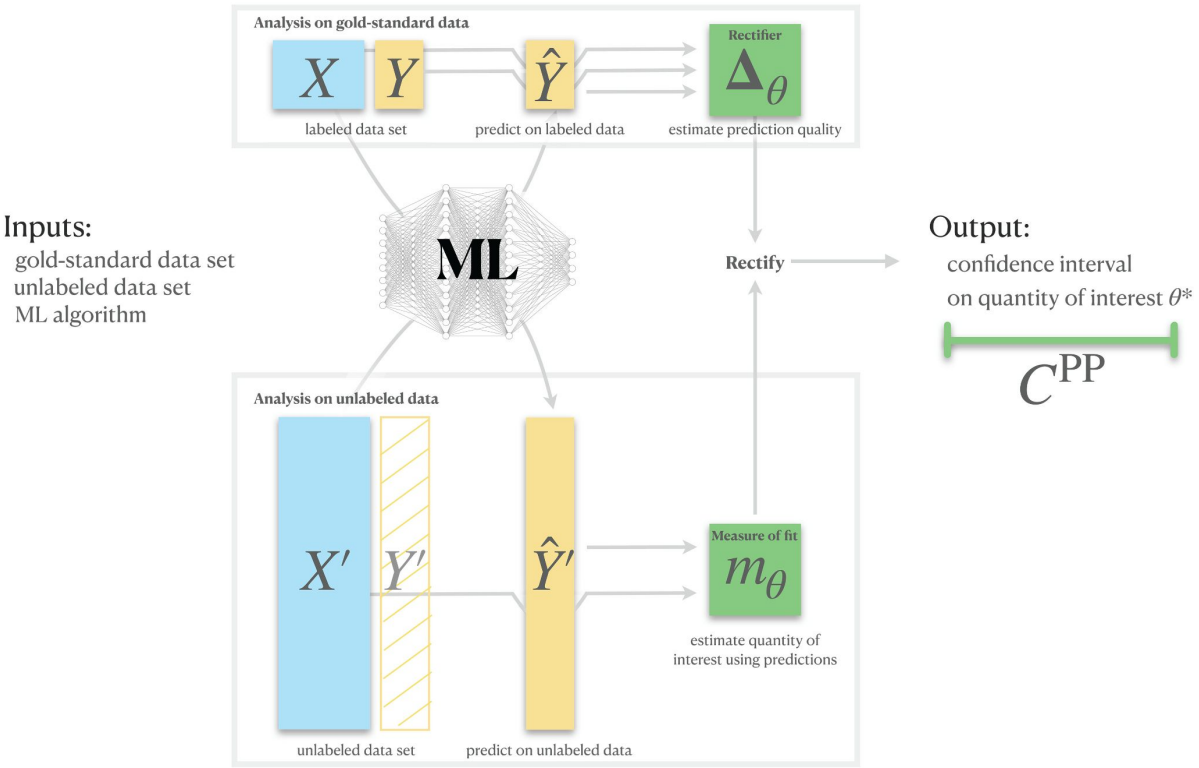
Estimates less variable, CIs too narrow

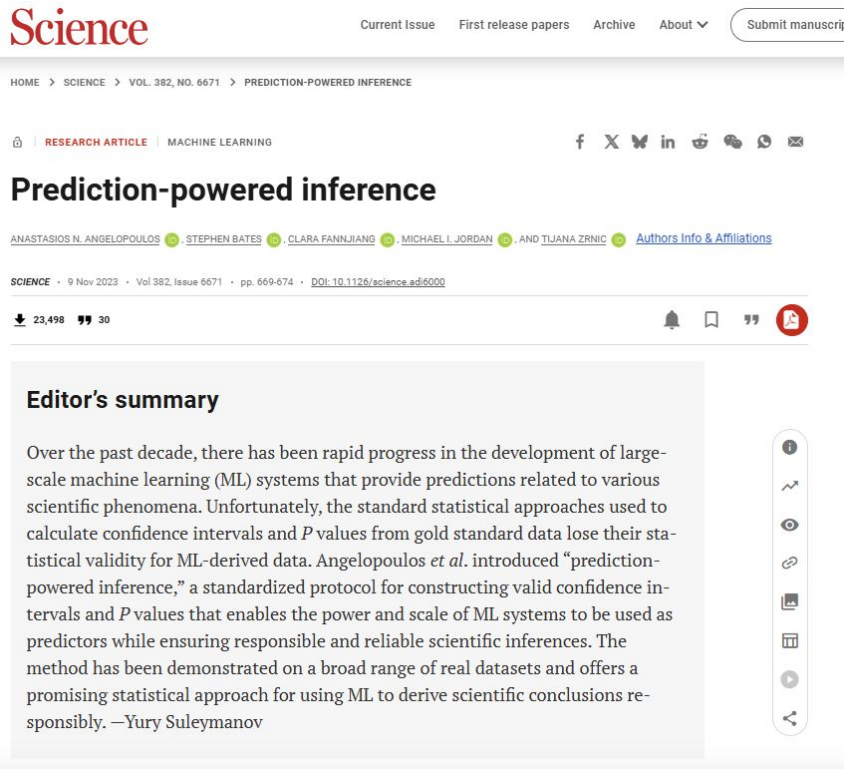
Estimates more variable, CIs wider

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

Prediction-Powered Inference (Angelopoulos et al., 2023)

'Rectify' the estimates and inference by modeling $Y - \hat{Y} = X_1 + X_2 + \ldots$





https://www.science.org/doi/10.1126/science.adi6000

Methods for inference with predicted data are rapidly evolving

Driven by the pace of AI/ML method development and the need for domain-specific solutions

The {ipd} R Package

Implements recent IPD methods in a consistent manner and with ‘tidy’ helper functions

Software Note

R Package

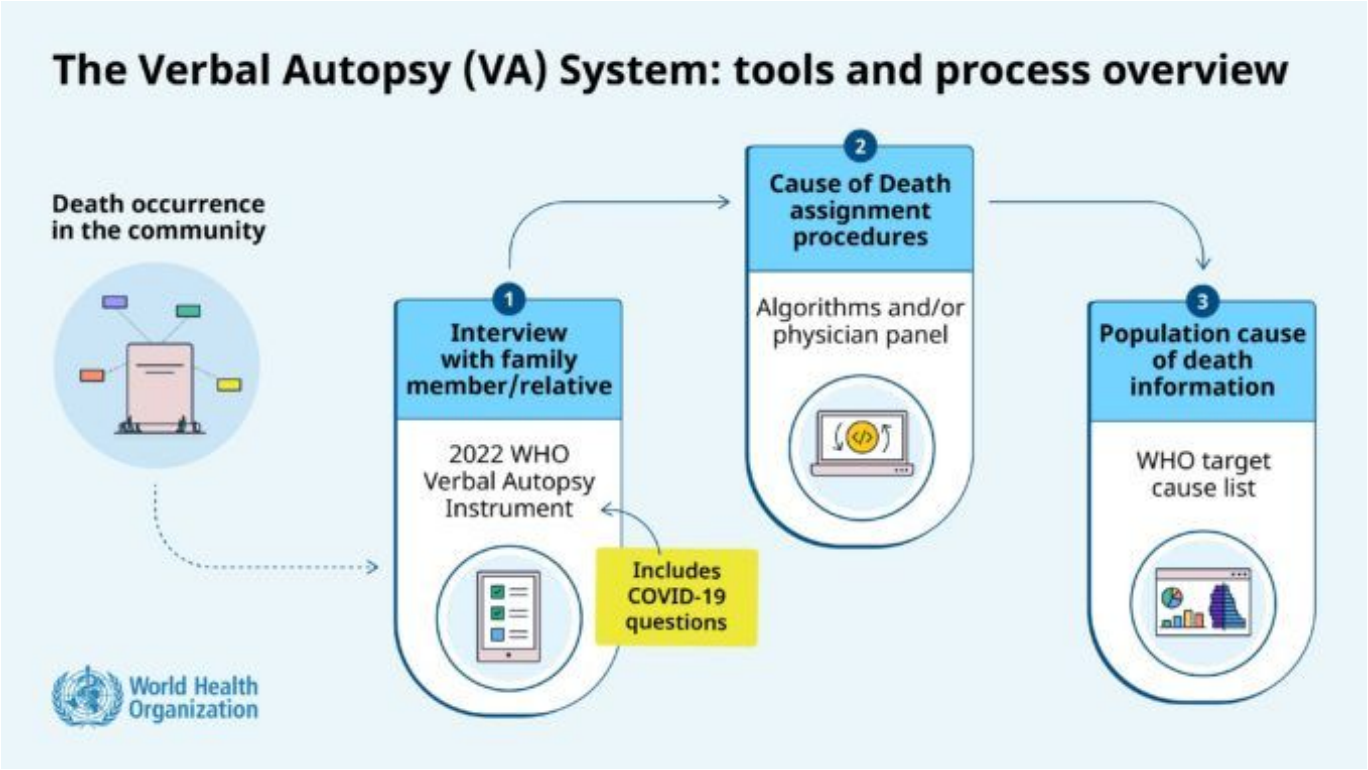
```
library(ipd)

df <- simdat(model="ols")|>

  filter(set_label != "training")

fit <- ipd(Y - f ~ X1,
           method = "chen",
           model = "ols",
           data = df,
           label = "set_label")

summary(fit)

results <- tidy(fit)
```

Case Study: Verbal Autopsy

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

< 1/3 of global deaths are assigned a medically-certified cause

Verbal autopsy (VA) is a time-consuming, resource-intensive process



https://publichealthnotes.com/verbal-autopsy-and-mpdsr/

Recent work leverages unstructured interviews for automated VA

Narrative summaries + AI/ML predictions would be less resource intensive than current standard

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 vision. 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']

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0308452

RecID: Symptoms from questionnaire	Extracted symptoms from textual data
A1: fever, cough, breath, chest pain, whooping	cough, whooping, cough, screaming, vomited, hot, swollen
A2: cancer, fever, sweating, breath low, yellow, abdomen pain, urine, hair, weight loss	stomach, vaginal bleeding, womb, cancer treatment, painful,stomach bleeding, urinating, swollen breast
A3: hypertension,asthma,low breath, weight loss, swollen legs, alcohol intake	passing stools, swollen
A4: hypertension, skin, urine, weight loss, swollen legs, excessive drinking	develop, sore toe, toes infection, wounds, grow infection, smell, worms, swollen
A5: fever, breath, chest pain, weight loss, excessive drinking, sweating	carrying, heavy box, twisted, pains, toes, affected, toes, swell, pain, blood, clot, removed, swell, huge, lump bursts, holes lump, chest, armpits, groin, throat, lumps, busted, puss coming
A6: TB, epilepsy, fever, cough, breath low, chest pain, diarrhoea, convulsion, vomiting, stiff neck	swollen, diarrhoea, stools, cough, go, toilet
A7: nil	pain, treated, looking better, weak, admitted, drips injection, worse
A8: hiv, fever, cough, low breath, yellow, alcohol, diarrhoea, vomit, weight loss, abdomen, swelling	sweats, lose weight, vomit, weak, hiv, water drips, diarrhoea, hallucinating, swollen stomach
A9: sweating, cough, low breath, urine, weight loss, paralysed, alcohol	cough, heavily, rot, worms, rotting, spread, thigh, buttocks, waist, extent, intestines, visible, abdomen
A10: fever, cough, vomit, weight loss, alcohol	cough, signs, vomits, low, energy, cough critical, tried, natural, fever
A11: hypertension, fever, low breath	fell, forehead, improve, speak, tears, rolled eyes, mixed blood, accident, urinate
A12: fever, cough, low breath, abdomen, swelling, headache, injury	initially, swollen ears, swollen, abdomen, critical, puss, ears, fever, cough
A13: nil	swollen, wounds, top, feet, wounds, burst, puss, swollen, feet, cuts, razor blade, rubbed
A14: fever, low breath, chest pain, diarrhoea, vomit, weight loss, hair,eyes sunken	committing, thinner, diarrhoea, oral, dehydration, solution, vomit
A15: chronic, fever, skin, weight loss, swollen legs, alcohol, smoking	pain, toe, long, kneel, cut, flesh wound, toe, swollen toe, infected
A16: fever, cough, diarrhoea, blood diarrhoea, vomit, weight loss	diarrhoea, vomit, weight loss, oral, dehydration, thrush
A17: hiv, stroke, low breath, weight loss, alcohol, chest pain	fell, weak, touch, hold, feeling, pains, arv, years, clear, pass, stools, giving, arv, sweats, minutes

https://doi.org/10.1371/journal.pone.0308452.t002

We studied how different AI/ML methods classify cause of death

Comparing KNN, SVM, BERT, and GPT-4

OpenReview.net Search OpenReview... Login

← Go to **COLM 2024 Conference** homepage

From Narratives to Numbers: Valid Inference Using Language Model Predictions from Verbal Autopsies

Shuxian Fan, Adam Visokay, Kentaro Hoffman, Stephen Salerno, Li Liu, Jeffrey T. Leek, Tyler McCormick

Published: 10 Jul 2024, Last Modified: 25 Aug 2024 COLM Everyone Revisions BibTeX CC BY 4.0

Research Area: Data, Science of LMs, Inference algorithms for LMs, LMs for everyone
Keywords: multiclass, inference, transportability, public health, verbal autopsy

TL;DR: We show how to perform valid statistical inference when using NLP predictions instead of directly observed data.

Abstract:
In settings where most deaths occur outside the healthcare system, verbal autopsies (VAs) are a common tool to monitor trends in causes of death (COD). VAs are interviews wit relative that are used to predict the decedent's COD. Turning VAs into actionable insights for researchers and policymakers requires two steps (i) predicting likely COD using the performing inference with predicted CODs (e.g. modeling the breakdown of causes by demographic factors using a sample of deaths). In this paper, we develop a method for va outcomes (in our case COD) predicted from free-form text using state-of-the-art NLP techniques. This method, which we call multiPPI++, extends recent work in "prediction-pow multinomial classification. We leverage a suite of NLP techniques for COD prediction and, through empirical analysis of VA data, we demonstrate the effectiveness of our approa transportability issues. multiPPI++ recovers ground truth estimates, regardless of which NLP model produced predictions and regardless of whether they were produced by a m GPT-4-32k or a less accurate predictor like KNN. Our findings demonstrate the practical importance of inference correction for public health decision-making and suggests that end goal, having a small amount of contextually relevant, high quality labeled data is essential regardless of the NLP algorithm.

Supplementary Material: zip
Code Of Ethics: I acknowledge that I and all co-authors of this work have read and commit to adhering to the COLM Code of Ethics on https://colmweb.org/CoE.html
Author Guide: I certify that this submission complies with the submission instructions as described on https://colmweb.org/AuthorGuide.html
Submission Number: 496

https://openreview.net/forum?id=QbCHIIqbDJ

GPT-4 Zero-Shot Prompt Used for COD Prediction

```
<narrative>
INPUT
</narrative>

<labels>
aids-tb: Patient died resulting from HIV-AIDs or Tuberculosis.
communicable: Patient died from a communicable disease such as pneumonia, diarrhea
or dysentery.
external: Patient died from external causes such as fires,
drowning, road traffic, falls, poisonous animals, suicide,
homicide, or other injuries.
maternal: Patient died from pregnancy or childbirth
including from severe bleeding, sepsis, pre-eclampsia and eclampsia.
non-communicable: Patient died from a non-communicable disease such as cirrhosis,
epilepsy, acute myocardial infarction, copd, renal failure, cancer, diabetes,
stroke, malaria, asthma.
unclassified: narrative does not contain enough information to predict cause of death.
</labels>

<options>
aids-tb,
communicable,
external,
maternal,
non-communicable,
unclassified
</options>

Which label from options best applies applies to the narrative?
If you are not sure, return your best guess.
Limit your response to one of the options exactly as it appears in the list.
```

Each narrative gets plugged in here

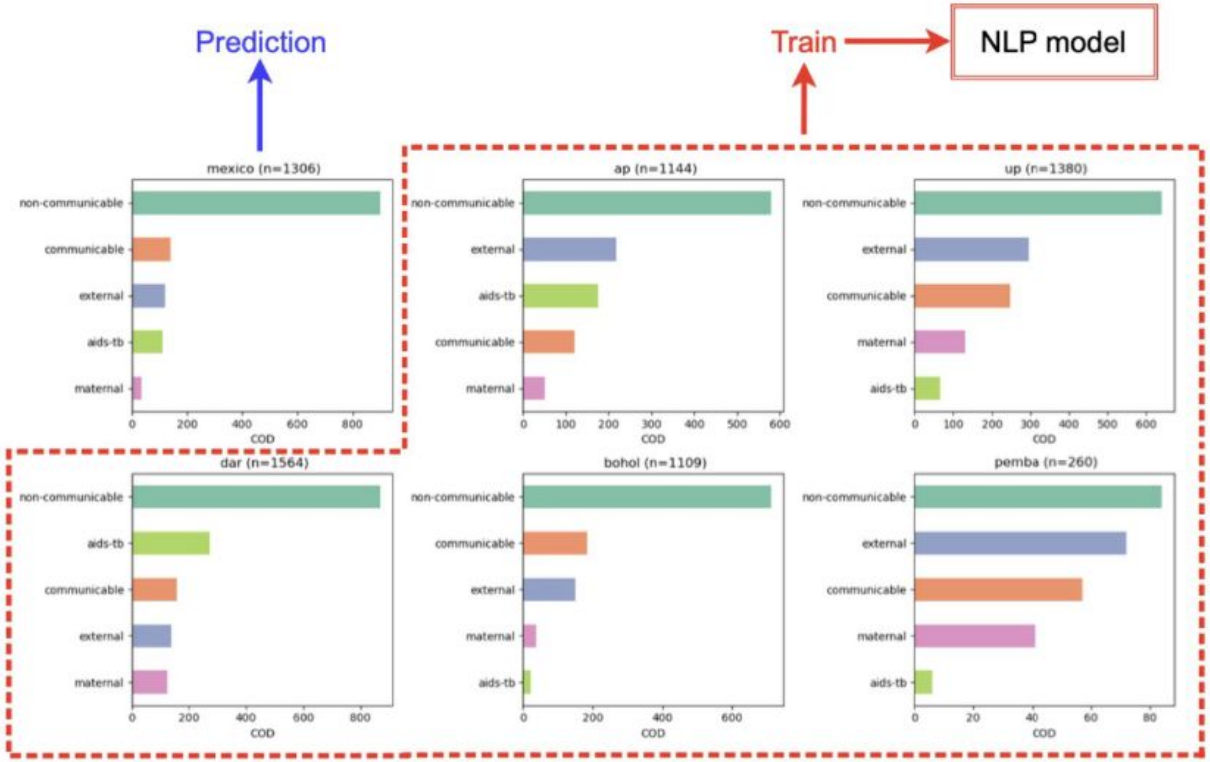
Context

Explicitly require output in this format

Instructions

Population Health Metrics Research Consortium

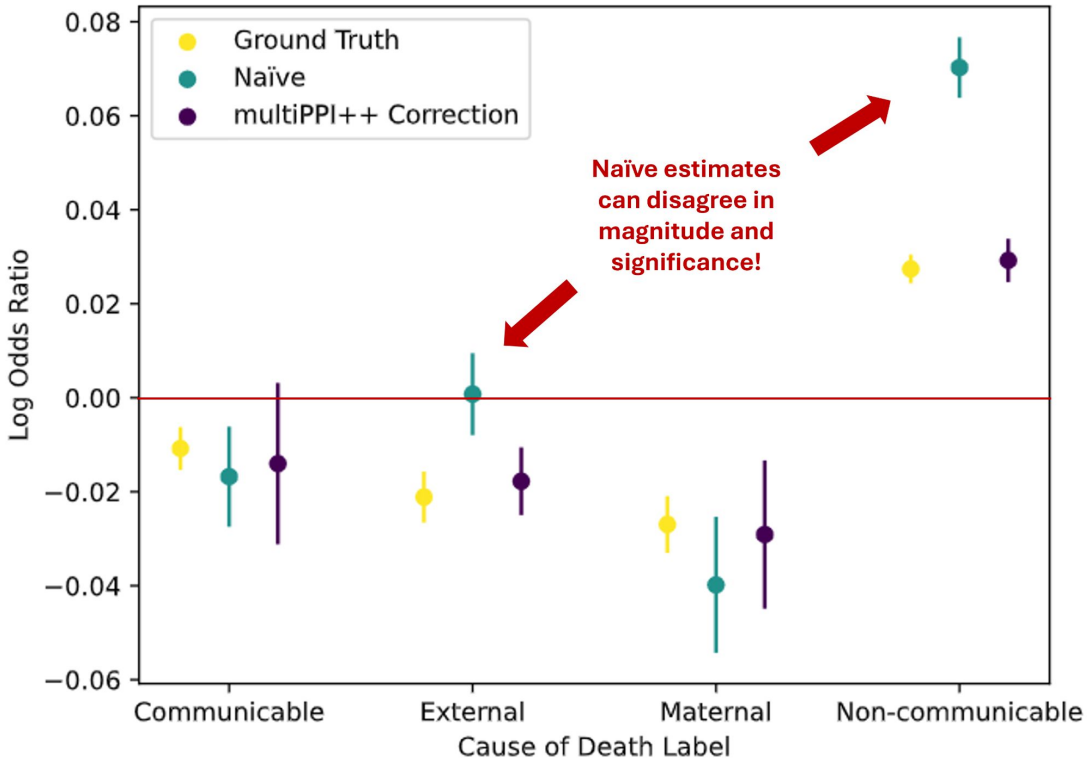
N = 6,763 observations from 6 sites in 4 countries; adults only, 5 cause of death labels



https://openreview.net/forum?id=QbCHIIqbDJ

And how IPD corrects inference on predicted cause of death

By extending an existing method (PPI++) for multiclass prediction



https://openreview.net/forum?id=QbCHIIqbDJ

Case Study: Body Mass Index

Adapted from *Post-prediction inference*. Source: jtleek.com/talks

‘Predictions’ don’t have to be complex to bias inference

The case of body mass index (BMI; kg/m^2) as an *imperfect* measure of adiposity

Three-Quarters of U.S. Adults Are Now Overweight or Obese

A sweeping new paper reveals the dramatic rise of obesity rates nationwide since 1990.

Listen to this article · 7:32 min Learn more Share full article 2.2K

Getty Images

By Nina Agrawal

Published Nov. 14, 2024 Updated Nov. 15, 2024, 2:17 a.m. ET

The paper defined “overweight” adults as those who were age 25 and over with a body mass index at or over 25, and “obese” adults as those with a B.M.I. at or over 30. The authors acknowledged that B.M.I. is an imperfect measure that may not capture variations in body structure across the population. But from a scientific perspective, experts said, B.M.I. is correlated with other measures of body fat and is a practical tool for studying it at a population level.

https://www.nytimes.com/2024/11/14/well/obesity-epidemic-america.html

An ‘algorithm’ is just a set of steps that map inputs to outputs

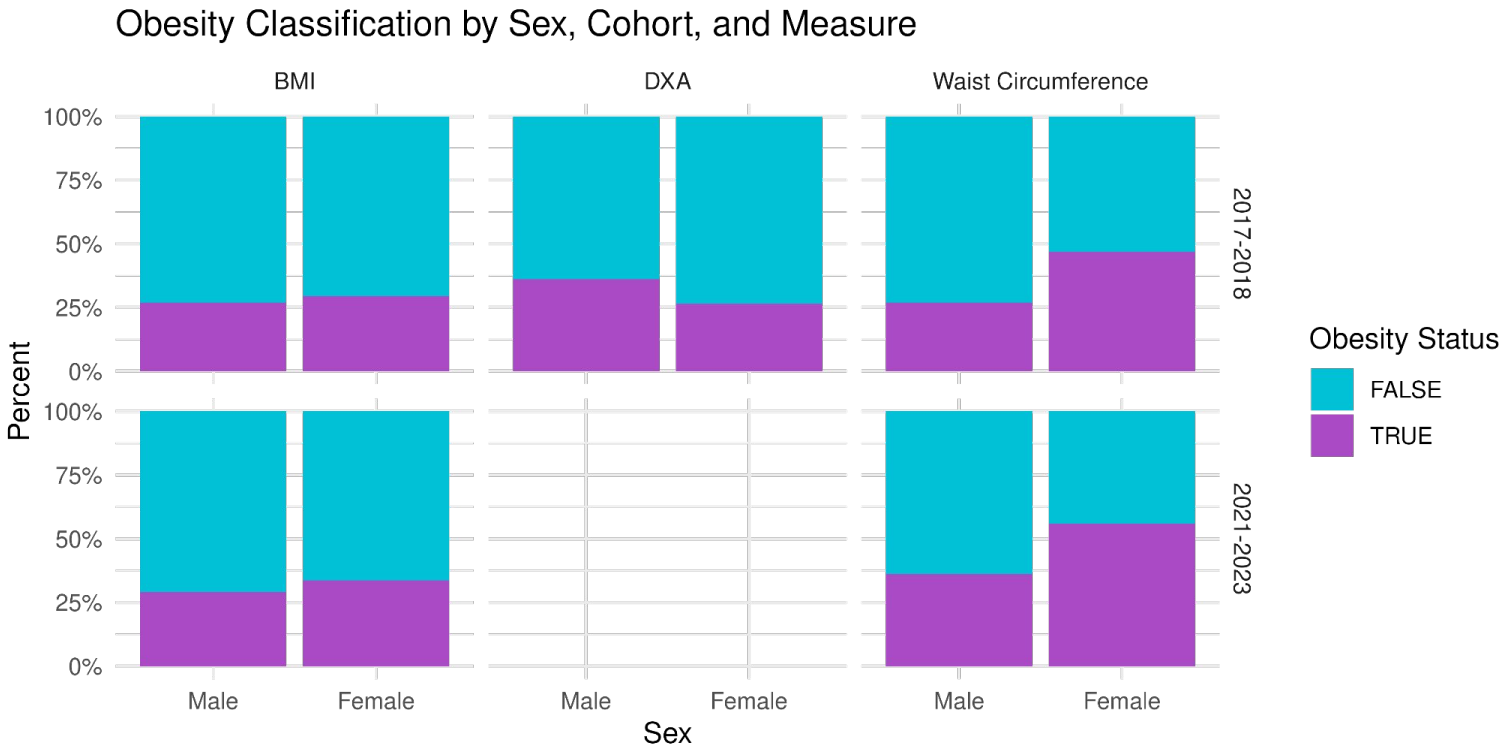
BMI is one of several potential adiposity prediction algorithms

- BMI (≥ 30 kg/m²)
- Waist circumference (men ≥ 102 cm; women ≥ 88 cm)
- DXA % body fat (> 30% men; > 42% women)

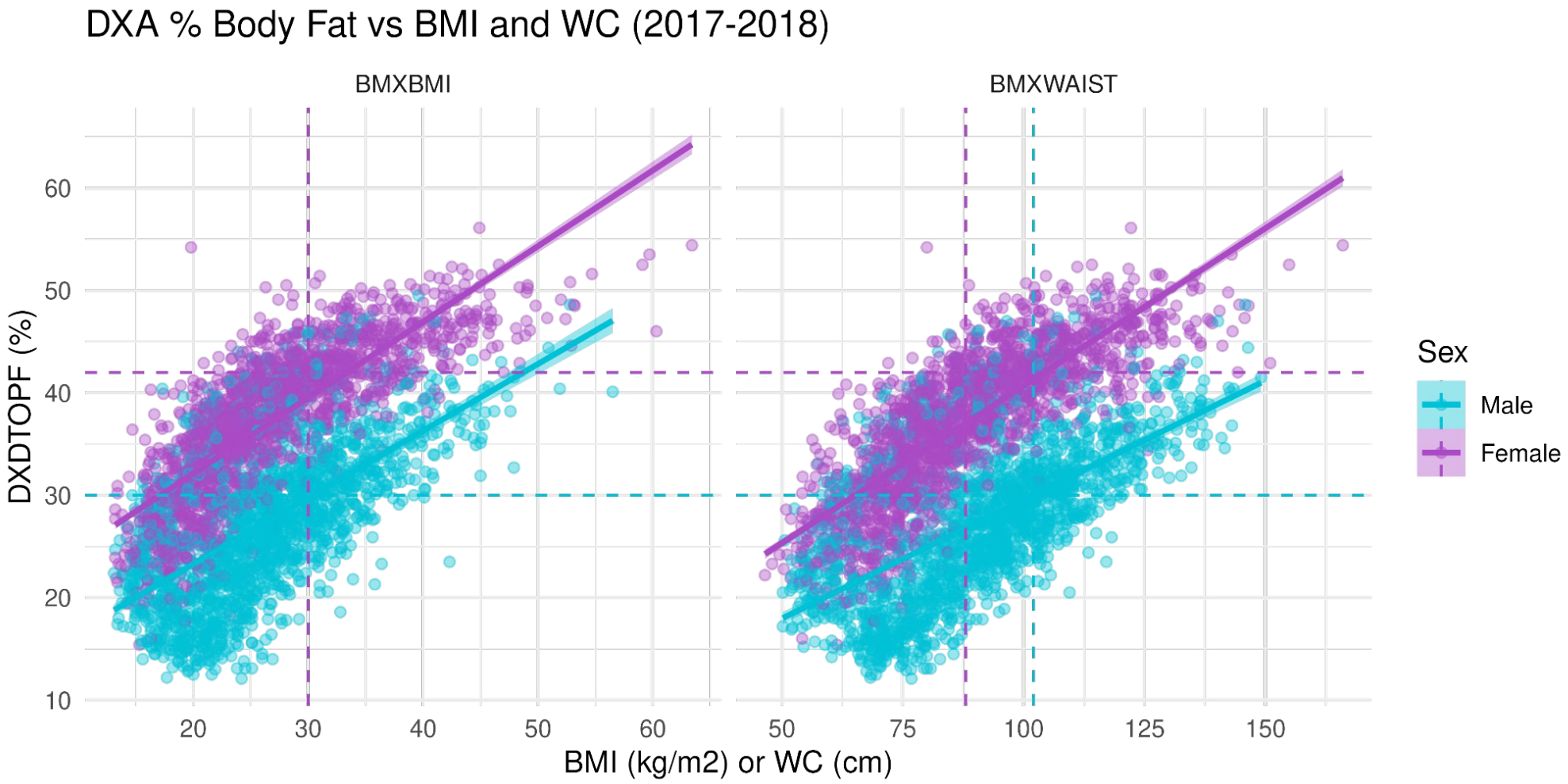
https://www.medrxiv.org/content/10.1101/2025.04.01.25325037v1.full.pdf

Using NHANES to understand population-level inference

BMI/WC/DXA all collected up until COVID-19, when DXA collection was suspended

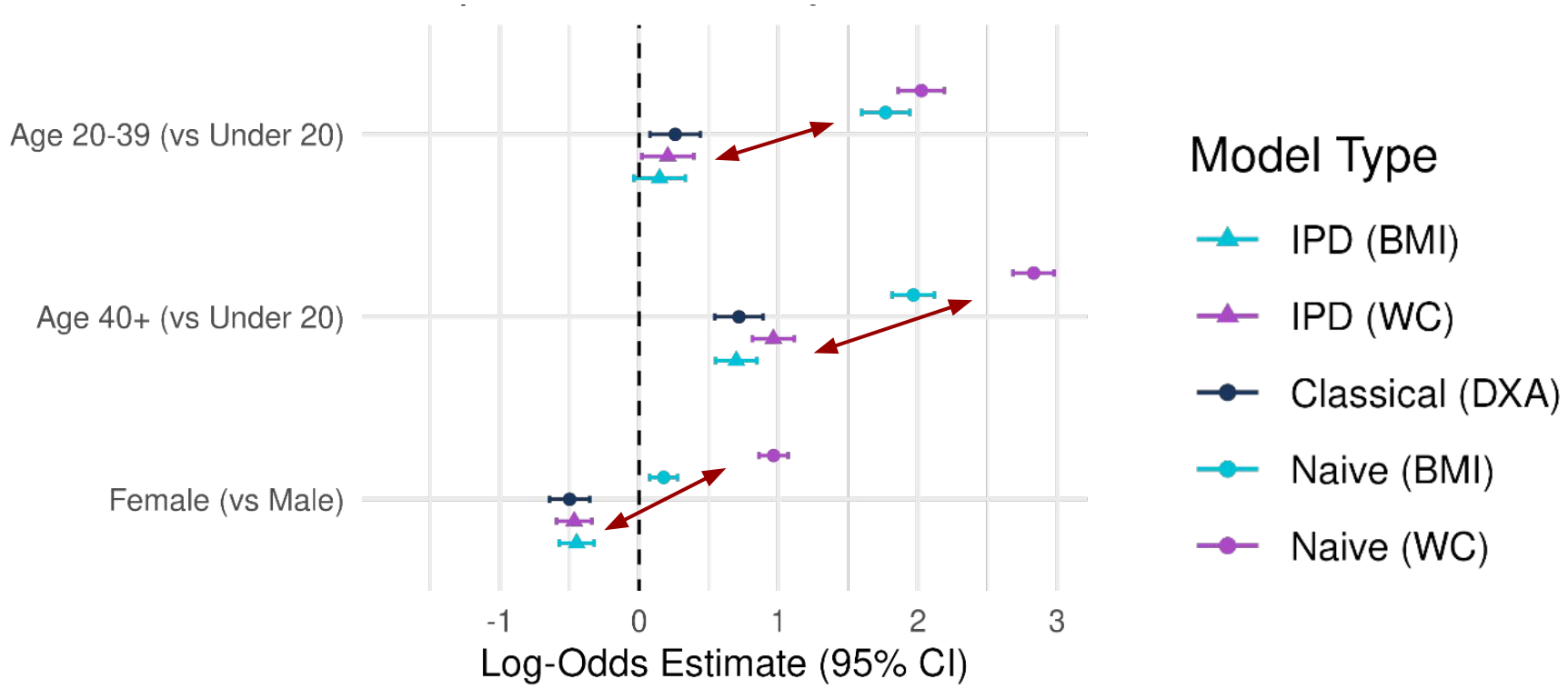


Mismatched quadrants suggest limitations in BMI/WC



Differences in sign and magnitude depending on the measure!

But IPD correction aligns with gold-standard DXA-based measure of obesity



Some final thoughts

This area is only becoming more relevant in this AI/ML era

- Increasing reliance on AI/ML raises questions about data quality/validity
- IPD is rapidly evolving, driven by need for rigorous methods to domain-specific problems
- We believe that open-source collaboration is the fastest way to success!

These Slides

ssalerno@fredhutch.org

https://github.com/ipd-tools

Fred Hutch
Cancer Center

Thank you!

https://mymodernmet.com/boston-dynamics-do-you-love-me-dance/