# Beyond AI Predictions: Valid Statistical Inference from Sparse Labels

[redacted]

## Abstract

Missing data creates a major challenge for statistical analysis because standard methods often depend on assumptions about why data are missing—assumptions that can't be tested. This paper evaluates Prediction-Powered Inference (PPI) and its cross-validated extension (Cross-PPI) as robust alternatives that leverage machine learning predictions while explicitly correcting for imputation error. Specifically, we focus on the application of these frameworks to medium-sized real-world datasets, an area that current literature does not extensively cover. Through theoretical analysis and empirical benchmarks on real-world datasets (Red Wine quality metrics and Census income classification), we demonstrate that both methods provide valid confidence intervals. Our results reveal three key insights: First, with a random forest imputer, PPI and Cross-PPI achieve nominal coverage while reducing interval widths by 30–40 percent compared to complete-case analysis. Second, when imputer quality degrades (e.g., using a shallow decision tree), Cross-PPI still maintains better coverage but requires wider intervals to do so. These findings highlight a tradeoff between imputer sophistication and inferential precision: while PPI methods outperform classical approaches even with weak imputers, their efficiency gains diminish sharply when predictions are poor. Nonetheless, the applicability of PPI to real-world datasets is strong. We provide practical guidelines for selecting imputers and labeled-data proportions to balance coverage and precision in resource-limited settings. Open-source implementations are included for immediate application.

## 1 Introduction

Missing values are intrinsic to modern data analysis, yet many researchers continue to utilize tools reliant on unverifiable assumptions. Classical procedures—complete-case analysis, likelihood-based estimation, and multiple imputation—rest on either *missing completely at random* (MCAR) or *missing at random* (MAR) conditions as defined by Rubin (10, 11). These convenient theoretical frameworks resist empirical validation from observed data alone, and even subtle violations—such as in *missing not at random* (MNAR) scenarios—can severely bias estimates and produce confidence intervals with inadequate coverage (7). These missing data assumptions cannot be validated precisely because the required information is, by definition, absent. This inherent paradox points to the need for alternative inferential approaches that don't rely on untestable assumptions about missing data mechanisms.

Recent AI advances have renewed interest in missing data imputation, suggesting promising new approaches. While large language models and advanced AI techniques offer powerful prediction capabilities, the fundamental distinction between accurate prediction and valid statistical inference remains critical. Naïve plug-in methods that fill in missing values with predictions and treat them as true ignore prediction uncertainty, often leading to unreliable statistical results. (14). Importantly, our investigation does not use generative AI to impute missing outcomes. Instead, we focus on uncertainty quantification in settings with partially observed outcomes — that is, where outcome labels are missing for a subset of the data. Importantly, our investigation does not employ generative AI to create imputed data, but rather examines principled statistical approaches to uncertainty quantification.

Many practical inference problems involve incomplete outcome data—where labels $Y$ are observed for only a small subset of covariate profiles $X$, due to cost, logistics, or privacy constraints. This setup can be naturally framed as a missing data problem, where the unobserved outcomes are treated as missing values. From a semi-supervised learning perspective, this corresponds to a dataset split into a labeled set $(X, Y)$ and a much larger unlabeled set $X$. Naïve methods that impute the missing $Y$s using predictions and treat them as observed can dramatically understate uncertainty. In contrast, we use prediction-powered inference to calibrate intervals using the labeled subset, explicitly accounting for prediction error in the unlabeled pool. This approach enables statistically valid inference under minimal assumptions about the missingness mechanism.

Prediction-Powered Inference (PPI) (1) introduces a framework that addresses this challenge by combining a small labeled subset with model predictions on a larger unlabeled pool. By using the average residual on labeled data to calibrate results, PPI offers a way to construct more reliable intervals by correcting for prediction error—without requiring strong assumptions about why data are missing. Its refinement, Cross-PPI (16), implements $K$-fold sample splitting to ensure that every prediction and residual is computed out-of-sample, thereby eliminating in-sample bias. Through explicitly accounting for prediction uncertainty, these methods enable valid inference where additional label collection is prohibitively expensive and classical missing-data assumptions remain suspect.

*Our principal research question is as follows:* **To what extent can prediction-powered inference methods (PPI and Cross-PPI) provide valid statistical inference with narrower confidence intervals compared to classical methods, without requiring strong missing data assumptions, when applied to moderate-sized scientific datasets across varying labeled proportions?** This question addresses a critical gap in the literature, as most PPI evaluations have focused on settings with advanced models or massive datasets. Instead, we investigate performance in contexts more representative of typical scientific research—datasets where each observation might represent an expensive experiment with human or animal subjects. By rigorously benchmarking these methods across varying labeled proportions, we explore whether prediction-powered approaches can provide the statistical validity of classical methods while achieving tighter confidence intervals than naïve imputation in resource-constrained scientific settings.

To investigate this question, we investigate two complementary analyses:

1. **Methodological benchmarking**. We evaluate PPI and Cross-PPI against three established baselines: (i)c̃lassical complete-case intervals; (ii)ñaïve plug-in intervals that ignore prediction error; and (iii)g̃old-standard intervals computed from the full set of labels when such labels are available.

2. **Robustness assessment**. Extensive Monte-Carlo experiments over a range of labeled-set sizes are performed on both a regression task (Wine Quality) and a classification task (Census Income), allowing us to examine interval width and stability under diverse conditions.

Taken together, the findings show that PPIand Cross-PPI consistently produce narrower and statistically confident intervals compared to classical analysis, even when the number of available labels is small.

## 2 Background and Literature Review

Traditional statistical inference relies on high-quality labeled data, yielding valid but often inefficient results with wide confidence intervals. However, gold-standard labels are typically expensive or scarce, making it difficult to scale such approaches in practice.

To address limited labeled data, researchers often turn to imputation methods, where machine learning predictions substitute for missing or unobserved values. A common approach is single imputation, using simple techniques like mean or regression-based substitution (15). While easy to implement, single imputation tends to underestimate uncertainty and distort relationships between variables, especially under missing not at random (MNAR) mechanisms. Regression-based imputation better preserves dependencies among variables but still introduces bias and underestimates variance. These approaches fail to offer statistically valid inference due to their deterministic nature.

Multiple imputation improves on this by generating several plausible datasets and averaging results to preserve variability (9). It allows for more valid statistical inference by accounting for uncertainty in the imputation process. However, it relies on strong modeling assumptions, is computationally demanding, and may struggle with complex, high-dimensional, or non-linear data. Inference can still be invalid if the imputation model is misspecified.

These limitations have motivated new inference procedures that directly adjust for prediction error in a model-agnostic way. As machine learning predictions are increasingly used in place of direct observations, there is a growing need to quantify and correct the bias they introduce (13).

Prediction Powered Inference (PPI), introduced by Angelopoulos et al. in 2023 (1), formalizes this idea. PPI combines sparse but trustworthy labels with abundant but imperfect predictions to produce unbiased and more efficient estimates of population parameters. It adjusts naive estimates using a rectifier that corrects for the average prediction error, yielding valid confidence intervals under minimal assumptions. A key advantage is that PPI works with any machine learning model, regardless of its internal structure.

Extensions like Cross-PPI (17) improve stability by using cross-fitting and ensemble predictions to reduce the variance from data splitting. PPI++ (3) further enhances efficiency and statistical power by optimizing over a tuning parameter $\lambda$. These methods have already seen real-world application, such as in clinical trials (8), where they enable reduced sample sizes without sacrificing validity.

Our study builds on this foundation, applying PPI and its variants to moderate-sized datasets where data collection is costly or constrained by ethical and practical concerns.

## 3 Preliminaries

This section presents the mathematical framework for Prediction-Powered Inference (PPI) and Cross-Prediction-Powered Inference (Cross-PPI) as introduced by Angelopoulos et al. (2) and Zrnic and Candès (17), respectively. We focus on the rectifier concept that enables valid statistical inference when using machine learning predictions.

### 3.1 Problem Setup

We consider a setting with a limited number of labeled data points $(X_i, Y_i)_{i=1}^{n}$ sampled i.i.d. from a distribution $P$, and a larger set of unlabeled data points $(\tilde{X}_i)_{i=1}^{N}$ with $N \gg n$. Our goal is to infer a population parameter $\theta^*$, such as $\mathbb{E}[Y]$.

### 3.2 Prediction-Powered Inference

PPI constructs the estimator:

$$\hat{\theta}^{\mathrm{PPI}} = \frac{1}{N} \sum\_i = 1^N f(\tilde{X} * i) - \delta \qquad (1)$$

where the rectifier $\delta$ is:

$$\delta = \frac{1}{n} \sum * i = 1^n (f(X\_i) - Y\_i) \qquad (2)$$

This correction yields an unbiased estimator of $\theta^*$ under weak assumptions:

$$\mathbb{E}\hat{\theta}^{\mathrm{PPI}}] = \mathbb{E}Y] = \theta$$

### 3.3 Cross-Prediction-Powered Inference

Cross-PPI extends PPI by training $K$ models using cross-fitting on labeled folds, then combining them. Predictions for each fold are made out-of-sample to avoid leakage.

$$\begin{aligned}\hat{Y}_i^{\text{unlabeled}} &= \frac{1}{K}\sum_{j=1}^{K} f^{(j)}(\tilde{X}_i) \\ \hat{Y}_i^{\text{labeled}} &= f^{(j)}(X_i) \qquad (5)\end{aligned}$$

The Cross-PPI estimator is:

$$\begin{aligned}\hat{\theta}^{\text{Cross}} = &\frac{1}{N}\sum_{i=1}^{N} \hat{Y}_i^{\text{unlabeled}} \\ &- \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i^{\text{labeled}} - Y_i\right) \qquad (6)\end{aligned}$$

### 3.4 Comparing PPI and Cross-PPI

- PPI requires a pre-trained model; Cross-PPI trains multiple models via cross-fitting.
- Cross-PPI uses labeled data more efficiently and improves stability.
- Cross-PPI's ensemble predictions often enhance accuracy for unlabeled data.

Both estimators remain unbiased and more efficient than classical approaches when $N \gg n$ and predictions are reliable.

### 3.5 Imputation Models

We use `RandomForestRegressor` from scikit-learn as our primary imputer—an ensemble method that reduces variance. We contrast this with a one-level decision tree ("stump"), which is highly interpretable but less accurate, to highlight the impact of model quality on PPI interval width.

## 4 Background of Study Datasets

This study relies on two benchmark datasets that differ in domain, scale, and data-generating process. The contrast between a controlled oenological laboratory study and a large-scale socio-economic survey allows us to test methodologies across both small/clean/regression and large/noisy/classification settings.

### 4.1 Red Wine Quality(UCI Machine Learning Repository)

The Red Wine Quality dataset was first released by Cortez *et al.* (4) and is hosted at the UCI Machine Learning Repository (5). It contains $n = 1{,}599$ observations, each describing a distinct Portuguese *Vinho Verde* red wine sample. For every sample, eleven physicochemical predictors are provided:

| Variable | Description |
|---|---|
| fixed_acidity | g tartaric acid / $\text{dm}^3$ |
| volatile_acidity | g acetic acid / $\text{dm}^3$ |
| citric_acid | g / $\text{dm}^3$ |
| residual_sugar | g / $\text{dm}^3$ |
| chlorides | g NaCl / $\text{dm}^3$ |
| free_sulfur_dioxide | mg / $\text{dm}^3$ |
| total_sulfur_dioxide | mg / $\text{dm}^3$ |
| density | g / $\text{cm}^3$ |
| pH | dimensionless |
| sulphates | g $K_2SO_4$ / $\text{dm}^3$ |
| alcohol | % (v/v) |

The target variable, `quality`, is an ordinal sensory score ranging from 0 to 10 assigned by expert tasters. Because `quality` is integer-valued and roughly bell-shaped (mean $\approx 5.64$, SD $\approx 0.81$), most studies frame the task as regression on the raw score. Key characteristics for our analysis:

- **Clean, nearly complete data**: there are no missing entries; preprocessing is limited to scaling/standardisation.
- **Small sample size**: with fewer than 2,000 rows, the dataset is well-suited for Monte-Carlo resampling and exhaustive cross-validation.
- **Low-noise laboratory measurements**: predictors are quantitative and collected under controlled conditions, making the data a prototypical example of a low-variance, tabular regression task.

### 4.2 Census Income Dataset

The Census Income dataset was extracted from the 1994 U.S. Census Bureau's Current Population Survey (CPS) and subsequently preprocessed for machine-learning applications by Kohavi and Becker (6). It comprises $n$ = 48,842 individuals, each characterized by fourteen demographic and employment variables, including: age, workclass, education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country. The predictive task is a binary classification problem: determining whether an individual's annual income exceeds $50,000. In this sample, approximately 24% of entries fall into the "`>50K`" category, yielding a moderate class imbalance.

Distinctive features:

- **Heterogeneous attribute types**: a mix of continuous (e.g. `age`, `hours-per-week`) and categorical variables with up to sixteen levels.
- **Real-world noise and missingness**: "?" marks missing entries in `workclass`, `occupation`, and `native-country`. Roughly 7% of rows contain at least one missing value.
- **Large-scale survey**: compared with the wine data, the Adult dataset is two orders of magnitude larger and reflects uncontrolled, population-level sampling.
- **Ethical considerations**: because the data encode sensitive attributes (race, sex, nationality), modelling must account for potential fairness and bias issues.

### 4.3 Why These Two?

Analysing both datasets side-by-side tests inference under contrasting conditions:

Table 1: Comparison of Wine Quality and Census Income datasets

| Characteristic | Wine Quality | Census Income |
|---|---|---|
| Domain | Chemistry / Sensory | Socio-economic |
| Task | Regression (ordinal) | Binary classification |
| $n$ | $1.6 \times 10^3$ | $4.9 \times 10^4$ |
| $p$ | 11 | 14 (mixed) |
| Missingness | None | $\sim$ 7% (MAR/MNAR) |
| Class imbalance | — | 24 % positive |
| Ethical risks | Low | Medium (fairness, privacy) |

This diversity allows us to benchmark our prediction-powered inference methods on both small-sample, low-noise data and large-sample, messy real-world data.

## 5 Methods

### 5.1 Gold-Standard Approach

In the ideal scenario where all data points are labeled, one can construct classical confidence intervals using standard statistical methods. In other words, in this context, our gold-standard estimators have access to all labels.

### 5.2 Classical Confidence Intervals for the Sample Mean

A classical example of interval estimation is for the population mean. The method for constructing a confidence interval for a sample mean (especially with unknown variance and small sample sizes) was introduced by William S. Gosset under the pseudonym "Student" (12). In his Biometrika paper, he derived the Student's $t$-distribution and established the textbook formula for a confidence interval on the mean of a normal distribution based on a finite sample. In this context, this gives us the labels-only mean.

### 5.3 Naïve Semi-Supervised Confidence Intervals (Treating Predictions as Truth)

In semi-supervised settings, where a dataset includes a small portion of labeled examples and a large portion of unlabeled examples, a straightforward but naive approach is to use a predictive model to label the unlabeled data and then treat those predicted values as if they were true outcomes when constructing a confidence interval. This approach ignores the model's prediction error and typically produces intervals with a biased centers and under-estimated standard errors, or intervals that are overly narrow and do not achieve nominal coverage. For a rigorous analysis of why this "plug-in" strategy fails and how to correct it, see Zhang et al. (14).

## 5.4 Prediction-Powered Inference (PPI)

With PPI, it becomes possible to use model predictions to improve the power of statistical inference while maintaining validity. Angelopoulos et al. (1) formally presented this approach in "Prediction-Powered Inference," showing how a small labeled dataset together with a large set of unlabeled examples (augmented with model predictions) can yield valid confidence intervals that shrink as the model's accuracy improves. This method makes no assumptions about the predictive model and guarantees nominal coverage by adjusting for prediction error.

## 5.5 Cross-Prediction-Powered Inference (Cross-PPI)

Cross-Prediction-Powered Inference adapts the original PPI idea by using a simple cross-validation (or sample-splitting) scheme. Standard PPI assumes you already have a pretrained predictor, which is often trained on a separate dataset. If you need to train your model on your current labeled set, then two subsets- one for training the predictor, another for estimating the rectifier- is required. Cross-PPI avoids this inefficient split by using cross-fitting: every labeled example is used for both training (in K-1 folds) and bias estimation (in its own fold). This leads to higher statistical efficiency, or tighter intervals with more stable estimates. Cross-PPI averaging over many splits and therefore reducing variability is particularly valuable in contexts where labeled datasets are small and the prediction model is imperfect (16).

## 5.6 Evaluation Procedure of Red Wine Dataset

**Experimental set-up.** Let $Y$ denote the 1 599 integer-valued quality scores and $X \in \mathbb{R}^{1\,599\times 11}$ the corresponding physicochemical covariates of the Wine Quality – Red data(5). We regard the full dataset as the population and fix the true mean $\theta^* = \frac{1}{1\,599}\sum_{i=1}^{1\,599} Y_i$ for subsequent coverage checks. To emulate scarce "gold-standard" labels, the index set $\{1,\ldots,1\,599\}$ is randomly permuted and split into a labelled portion $\mathcal{L}_n$ of size $n \in \{200,\ 400,\ 800\}$ and an unlabelled portion $\mathcal{U}_n$ of size $N = 1\,599 - n$. All experiments are repeated independently $T = 100$ times for each $n$.

**Predictive model.** Throughout, the regression function $f(\,\cdot\,)$ is taken to be a random-forest regressor with 100 trees and default SCIKIT-LEARN hyper-parameters. For Cross-PPI we further impose a three-fold (non-overlapping) partition of $\mathcal{L}_n$ and fit fold-specific forests $f^{(1)}, f^{(2)}, f^{(3)}$.

**Estimators compared.** Five methods are evaluated: (i) a "gold-standard" estimator that has access to all labels; (ii) the classical labels-only mean; (iii) a naïve semi-supervised mean that treats forest predictions on $\mathcal{U}_n$ as ground truth; (iv) in-sample PPI, which corrects the naïve mean by the in-sample residual average on $\mathcal{L}_n$; and (v) Cross-PPI, which instead uses out-of-sample residuals obtained via the three-fold cross-validation scheme.

**Confidence-interval construction.** For each replicate and each estimator $\widehat{\theta}$ we form a two-sided confidence interval $[\widehat{\theta} - h,\ \widehat{\theta} + h]$ with nominal coverage 90% ($\alpha = 0.10$), where $h = 1.645\,\widehat{\mathrm{SE}}(\widehat{\theta})$. The PPI standard-error formulas follow the variance decompositions in Angelopoulos *et al* (1).

**Performance metrics.** Our evaluation approach differs between the two datasets. For the Wine Quality dataset, which contains no missing values, we compare our methods against a gold standard derived from the complete dataset. For every $(n, t)$ pair, we record the interval width $w = 2h$ and assess how closely PPI and Cross-PPI intervals align with gold standard intervals. This serves as a sanity check, ensuring our methods produce results similar to what would be obtained with fully labeled data.

By aggregating across $T = 100$ replicates, we calculate the mean width $\overline{w}(n) = \frac{1}{T}\sum_{t=1}^{T} w_{n,t}$ for each method and compare these against the gold standard width. This demonstrates whether PPI methods can achieve precision comparable to the gold standard while using substantially fewer labeled examples.

For the Census Income dataset, which contains approximately 7% missing values, we cannot establish a reliable gold standard for comparison. Instead, we focus solely on comparing the relative performance of different methods in terms of interval width and stability across replicates, without making claims about coverage of a true parameter.

**Coverage Definition.** Our evaluation relies on the concept of *coverage*, which refers to how often a confidence interval contains a fixed reference value $\theta^*$ across repeated trials. For each method and sample size, we compute empirical coverage as

$$\hat{C}(n) = \frac{1}{T}\sum_{t=1}^{T} I_{n,t},$$

where $I_{n,t}$ is an indicator for whether the confidence interval in trial $t$ (with $n$ labeled points) contains $\theta^*$, and $T = 100$ is the number of Monte Carlo replicates.

For the *Wine Quality* dataset, which contains no missing values, we define our reference value as

$$\theta^* = \frac{1}{1,599}\sum_{i=1}^{1{,}599} Y_i \approx 5.636,$$

the average wine quality score across the full dataset. While this is not necessarily the true population mean, it serves as a fixed benchmark for assessing interval coverage within this dataset.

In contrast, we do not report coverage for the *Census Income* dataset. Because roughly 7% of its rows have missing entries and the mechanism behind this missingness is unknown, we cannot reliably define a "true" value for the population proportion earning above $50K. Any value calculated from the observed data could be biased due to non-random missingness. Without a valid reference point, coverage cannot be meaningfully assessed. Instead, we evaluate interval width and trends in performance qualitatively.

This approach allows us to compare methods in settings with and without missing data. High empirical coverage (close to the nominal 90%) indicates reliable inference, while undercoverage suggests intervals that are too narrow and fail to reflect uncertainty properly.

## 5.7 Simulation

Simulation is not our primary preoccuption in this paper. However, by simulating data, we evaluate the benefits of the proposed PPI and Cross-PPI frameworks when true parameters, such as a mean or regression coefficient, are known. This supplements our investigations of PPI and its performance when applied to analyses of medium-sized, real-world datasets. As in previous sections, we compare the performance of PPI estimates and the confidence intervals with labeled-only results and imputed results.

**Design.** Our code generates synthetic data with known parameters (true mean $\mu = 5$, regression coefficient $\beta_1 = 2$), then compares the following four estimators:

1. **Labeled-only** (standard supervised estimation)
2. **Naive imputation** (treating ML predictions as observed outcomes)
3. **PPI** (bias-corrected using in-sample residuals)
4. **Cross-PPI** (bias-corrected using out-of-sample CV residuals)

This is done using two models of differing complexity and two datasets described previously:

- A small labeled dataset (n=50) and large unlabeled dataset (N=800)
- Two ML models: random forests (RF) and shallow decision trees (max depth =2)

**Example: Mean Estimators.** Using standard PPI, for the mean, we use:

$$\hat{\mu}_{\text{PPI}} = \underbrace{\frac{1}{N}\sum_{j=1}^{N} f(X_j)}_{\text{Imputation}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))}_{\text{Rectifier}} \quad (7)$$

Next, using weighed least squares on the labeled and unlabeled datasets with weights $1/n$ and $1/N$ respectively, we use out of sample predictions from K-fold CV to build:

$$\hat{\mu}_{\text{Cross-PPI}} = \frac{1}{N}\sum_{j=1}^{N} f(X_j) + \frac{1}{n}\sum_{i=1}^{n}(Y_i - f^{\text{CV}}(X_i)) \quad (8)$$

**Implementation.** The code achieves three main objectives. First, it generates data from the linear model:

$$Y = 5 + 2X_1 + X_2 + X_3 + \epsilon, \quad \epsilon \sim N(0,1) \quad (9)$$

with covariates $X\ N(0, I_3)$.

Then, to train our models, we

- Fit RF and shallow trees to labeled data
- Compute normal and CV predictions
- Implement both standard and cross-validated PPI

We then calculated point estimates and 95 percent confidence intervals for the true mean (detailed prior) and the first regression coefficient, using robust standard errors.

# 6 Results

## 6.1 Red Wine Results

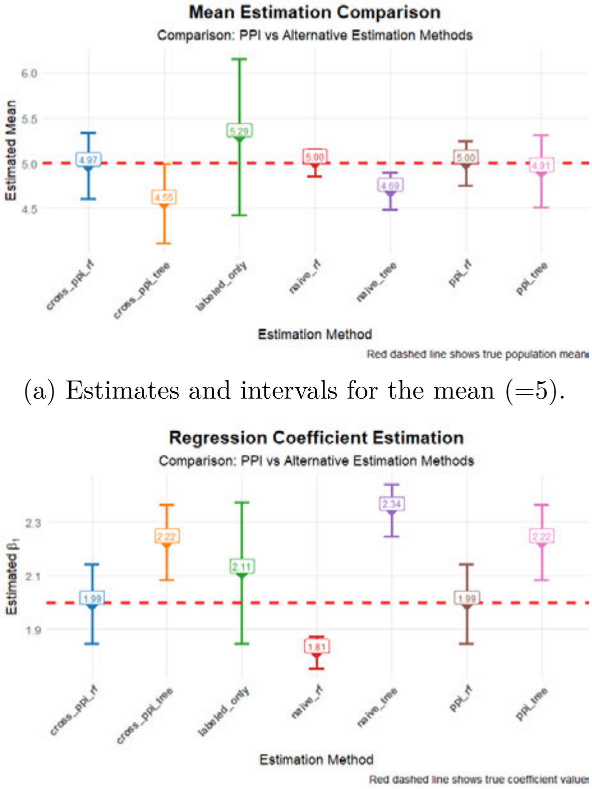
Figure 1b shows that, with the random-forest imputer, all methods achieve empirical coverage very close to the nominal 90 % (panel a) and produce relatively narrow intervals (panel b). By contrast, when a one-level tree is used, coverage falls below 0.90—especially at $n = 200$—and average CI widths increase substantially. The grouped-interval plots in Figure 2 and the bar charts at $n = 800$ in Figure 3 further illustrate that downgrading to a shallow tree roughly doubles the width of PPI and Cross-PPI intervals, pulling them nearer to the classical labels-only benchmark.

## 6.2 Census Results

As shown in Figure 4 (panels a and b), average CI widths for both PPI and Cross-PPI decrease with increasing label budget $n$, but this contraction is substantially slower when using the one-level tree compared to the random-forest imputer. At the largest budget ($n = 800$), Figure 5 confirms that the shallow tree pushes PPI and Cross-PPI widths much closer to the classical benchmark. Finally, Figure 6 contrasts all 100 replicate intervals side–by–side: under the weaker imputer, intervals not only become longer on average but also show greater dispersion, underscoring how inference precision hinges on imputer quality.

## 6.3 Simulation Results

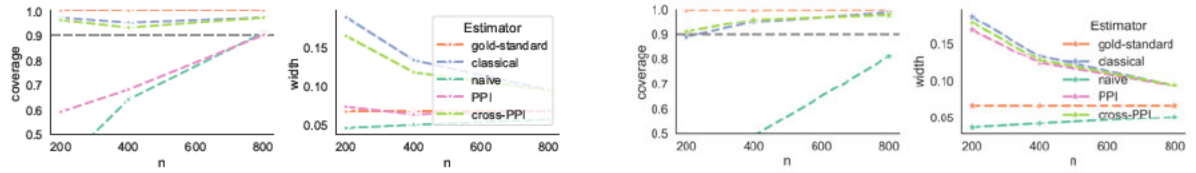
The following visualizations depict both the estimates and the widths of the associated intervals using each of the itemized methods described in the Methods section:



(a) Estimates and intervals for the mean (=5).

(b) Estimates and intervals for the first regression coefficient (=2).

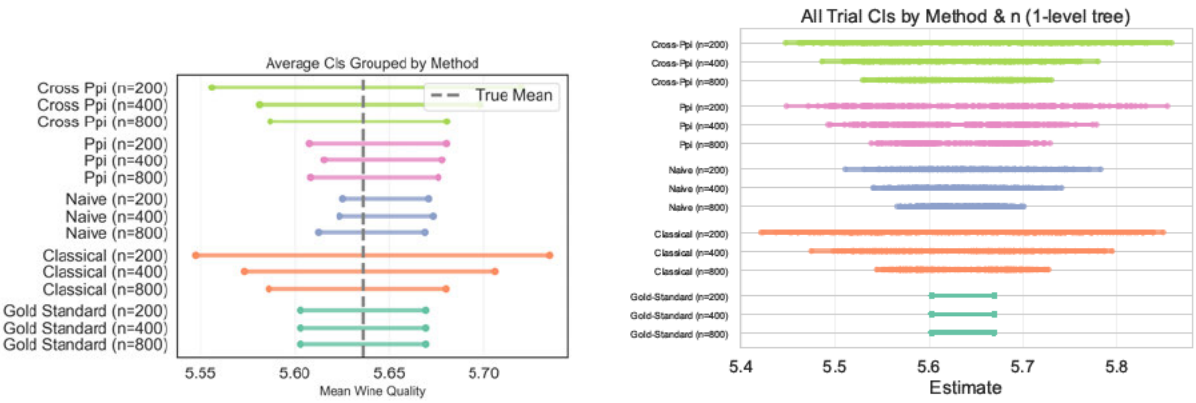
Figure 7: PPI frameworks display increases in efficiency compared to labeled-only approaches, and appropriate handling of error compared to naive imputation approaches.



(a) Random-forest imputer

(b) One-level tree imputer

Figure 1: Empirical 90 % coverage (left) and average CI width (right) for the red–wine mean under two imputers. The random forest achieves tighter intervals and nominal coverage, whereas the one-level tree yields wider intervals.



(a) Random-forest imputer

(b) One-level tree imputer

Figure 2: Mean 90 % confidence intervals for the red–wine mean, grouped by estimator and label budget. The one-level tree produces noticeably wider intervals than the random forest, especially at small $n$.

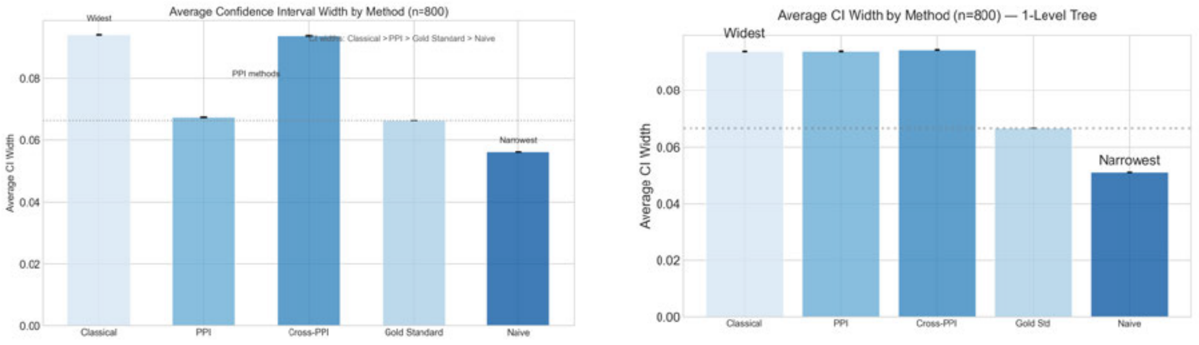
The true mean and true first regression coefficient are indicated with a dotted red horizontal line in both figures. Briefly, in both cases-mean estimation and coefficient estimation-the PPI methods (both Cross-PPI and standard PPI) outperform naive or labeled-only techniques, as the underlying theory would indicate. Specifically, in the case of mean estimation, PPI intervals were more stable than naive intervals, capturing the true mean in the interval even as the underlying ML model changed. The more complex model, the random forest, had tighter intervals than its counterpart, the shallow tree, both within naive results and PPI results. PPI intervals were wider than the naive intervals, holding the underlying ML model constant. Both were narrower than the labeled-only interval. This hierarchy:

$$\mathrm{Width}_{\mathrm{Gold\ standard}} < \mathrm{Width}_{\mathrm{PPI}} < \mathrm{Width}_{\mathrm{Labeled}} \tag{10}$$

reflects how PPI gain efficiency over labeled-only approaches, but represent proper uncertainty quantification compared to imputation methods.

The same patterns can be seen in the results pertaining to regression coefficient estimates. Naive intervals are narrow, and fail to capture the true parameter as a result. The PPI intervals fare better: their width allows them to estimate the true coefficient value better. This effect is stronger when using a more complex model, specifically the random forest rather than the shallow tree.

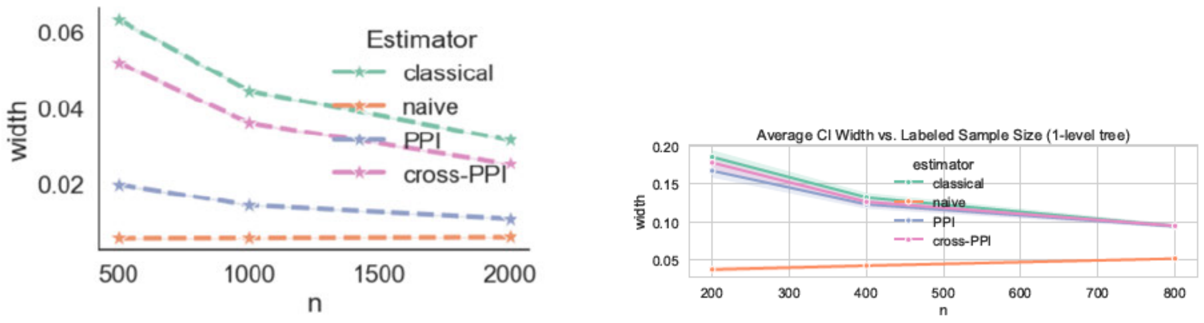
The purpose of our paper is not simply to evaluate or validate the accuracy of PPI, but to detail how the framework can be applied to restricted, medium-sized datasets. Further distinctions between standard PPI, cross-PPI, and their performance are detailed in the following sections as they pertain to real-world datasets.



(a) Random-forest imputer

(b) One-level tree imputer

Figure 3: Average CI widths across imputation methods with 800 labeled samples using Red Wine Data set.(a) Random forest imputer: PPI methods yield narrower CIs than Classical, with Naive narrowest.(b) One-level tree imputer: CI widths for PPI methods increase, approaching Classical levels.



(a) Random-forest imputer

(b) One-level tree imputer

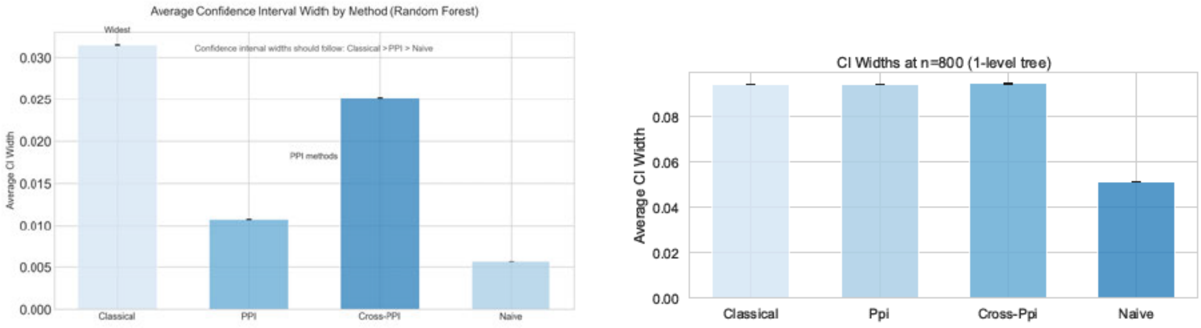
Figure 4: Average CI width vs. label budget $n$ for two imputers using red wine data. Prediction-powered intervals widen markedly when the imputer is downgraded from a random forest to a single-split tree.

## 7 Discussion

The empirical results across both data sets reveal a clear and intuitive ordering of interval precision. For every label budget we examined, classical complete-case intervals were widest, followed by PPI / Cross-PPI, then (when available) gold-standard intervals, with naïve plug-in intervals narrowest. This hierarchy, visible in the width–versus–$n$ curves of Figure 4 and the grouped bars in Figures 3–5, is explained because classical methods ignore the unlabeled pool and therefore pay a $1/n$ variance penalty; PPI methods use the unlabeled predictions but must still account for prediction error; a fully labeled ("gold") analysis exploits all $n+N$ responses; and the naïve approach attains the smallest nominal standard error only because it pretends the predictions are exact and therefore understates uncertainty.

**Simulation check.** The synthetic experiments in Figure 7 reinforce the same story. For both the mean and the regression-coefficient tasks, PPI and Cross-PPI capture the true parameter far more often than naïve plug-in, while still producing intervals much narrower than the labeled-only baseline. The random-forest model gives the tightest and most stable intervals; the shallow tree widens them and increases their spread—mirroring what we saw in the real data.

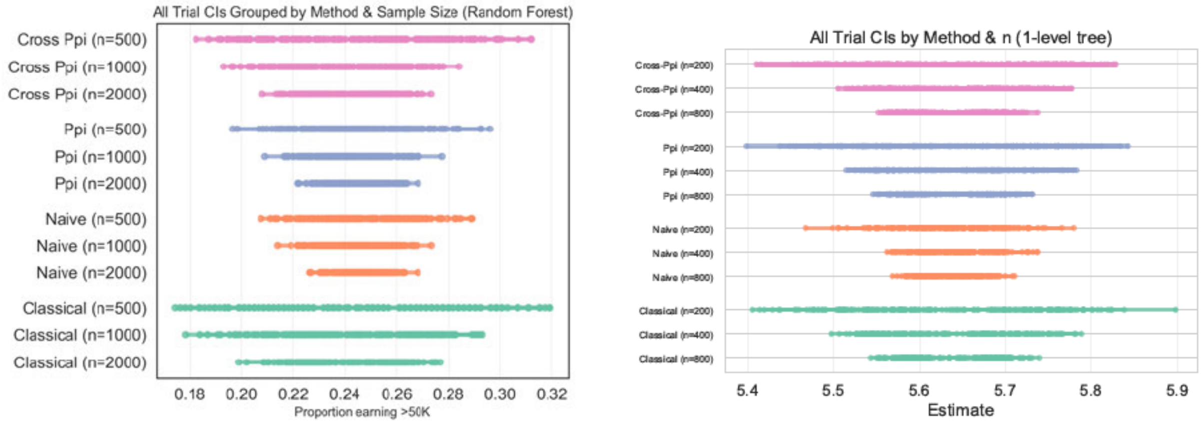
**Model complexity matters.** Replacing the random-forest imputer with a one-split stump roughly doubled the width of PPI and Cross-PPI intervals while leaving classical intervals unchanged (Figures 1 and 4). Because the stump produces far cruder predictions, its residual variance feeds directly into the PPI standard-error formula, inflating the interval.



(a) Random-forest imputer

(b) One-level tree imputer

Figure 5: Average CI width at $n = 800$ using the Cleveland Heart Dataset. The shallow tree pushes PPI and Cross-PPI intervals much closer to the classical benchmark.



(a) Random-forest imputer

(b) One-level tree imputer

Figure 6: Every CI from 100 Monte-Carlo replicates, grouped by estimator and label budget. Intervals lengthen and disperse with the weaker imputer, confirming PPI's sensitivity to prediction quality.

Conversely, the random forest's lower residual variance enables PPI to approach the gold-standard width on the Wine data and to remain far tighter than the classical benchmark on the Census task.

**Performance by sample size.** Interval widths shrink monotonically as the number of labels grows, but the gap between methods narrows more slowly (panel b of Figure 4a). At the smallest budget ($n = 200$ for Wine, $n = 500$ for Census) Cross-PPI intervals are about 30–40% narrower than classical ones when using the random forest; the advantage drops to roughly 10% when using the stump. Cross-PPI is consistently a few percentage points tighter than in-sample PPI, reflecting the variance reduction obtained by using strictly out-of-sample residuals.

**Task-specific highlights.** For the Wine Quality data, keeping only $n = 400$ labels and running PPI with a random-forest model produces a confidence interval that is almost the same width as the interval built from all 1,599 labels. (Figure 3a). With the stump, PPI widths move toward the classical level but remain meaningfully shorter, confirming that even a weak imputer is preferable to discarding the unlabeled data (Figure 3b).

For the Census Income classification task, no gold-standard benchmark is possible because $\sim 7\%$ of responses are genuinely missing. Nevertheless, the same width ordering shows. At $n = 2\,000$ labeled cases, Cross-PPI with the random forest cuts the classical width by roughly half (Figure 5a), whereas the stump reduces it by only $\approx 15\%$ (Figure 5b). Naïve plug-in intervals are always shortest, but their variance is understated and they should not be used for inference.

**Practical Implications.** Our study suggests several useful insights for statistical analysis. First, when few labels are available, prediction-powered inference methods produce narrower confidence intervals than classical analysis, especially when using random forest models rather than simple decision trees. Second, although the naïve plug-in intervals are the narrowest, our experiments indicate that they tend to fall short of the nominal 90 % coverage, making them less dependable for inference. Our results show that prediction-powered inference responds well to model quality: with random forests, it produces intervals almost as narrow as those from fully labeled data, while with simpler decision stumps, it produces wider intervals closer to classical methods—without needing strong assumptions about why data are missing.

Based on our Monte Carlo simulations and real-data evaluations, we suggest the following for researchers with limited labeled data. When very few labels are available (n < 0.05N), Cross-PPI provides a good balance between achieving close to 90% coverage and maintaining reasonably narrow intervals. With more labeled data (0.1N < n < 0.2N), Cross-PPI continues to perform well, offering narrower intervals than classical methods. We caution against using the standard PPI method, which can suffer from data leakage. Researchers should avoid using the naïve approach despite its seemingly narrow intervals—our coverage analysis shows it consistently falls below the target 90% rate. The most noticeable improvements in interval width occur when there are many unlabeled observations and the prediction model performs well, making these methods particularly helpful when gathering labeled data is costly but feature data is readily available.

**Limitations and Future Work** We have a couple limitations to mention as followed. We used random forest models with default parameters; other predictive models might further improve PPI performance. We also focused on simple estimands (means and proportions); extending to other parameters like quantiles or regression coefficients $\beta^*$ would be valuable future work.

Finally, PPI methods assume the labeled data $\{(X_i, Y_i)\}_{i\in\mathcal{L}_n}$ represents the same population as the unlabeled data $\{X_i\}_{i\in\mathcal{U}_n}$. Developing diagnostics to detect when this assumption fails would be helpful and interesting for researchers.

## 8 Conclusion

This study shows that PPI and Cross-PPI can deliver tighter confidence intervals than classical complete-case analysis while still acknowledging prediction error. When we intentionally weaken the imputer to a single-split stump, the intervals widen toward the classical benchmark yet remain shorter, confirming that even modest predictive signal is worth exploiting. Naïve plug-in intervals, though visually smallest, miss the nominal coverage far too often and are therefore unsafe for inference.

In sum, Prediction Powered Inference lets statisticians reach valid conclusions without restricting them to inefficient methodologies. It bridges the gap between traditional inference and modern machine learning by providing a way to exploit predictive information without sacrificing statistical rigor. As machine learning continues to grow and permeate the process of scientific inquiry, methods like PPI will be increasingly essential for ensuring that inferences drawn from predictions are trustworthy, reproducible, and valid.

PPI empowers statistical strategy by providing a way to reduce variance, control bias, and tighten confidence intervals around valid, consistent estimates. This is a departure from labeled-only inference or direct imputation, where unlabeled data is ignored or bias is uncontrolled respectively. In practical implementations of PPI, different choices of ML models will still yield valid results. More complex models, such as random forests, usually lead to better predictions.

Our contributions explore the application of PPI in the absence of large AI models, predictions, and datasets. When labels are limited or expensive, and when data is moderately-sized, our results demonstrate that the advantages of PPI still hold. To summarize, we found that our PPI intervals were narrower than labeled-only intervals, demonstrating efficiency gain. PPI intervals also remained wider than Naive results, which represents proper uncertainty quantification. Further research could focus on common obstacles like extremely small datasets (n<30), optimal hypertuning of underlying ML models, and cases where the unlabeled data distribution differs significantly from labeled data.

[redacted]

[redacted]

## References

[1] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671): 669–674, 2023. doi: 10.1126/science.adi6000.

[2] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[3] Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2024.

[4] Paulo Cortez, Antonio Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. doi: 10.1016/j.dss.2009.05.016.

[5] Dheeru Dua and Efi Karra Taniskidou. UCI machine learning repository. https://archive.ics.uci.edu/, 2019. Accessed May 20, 2025.

[6] Ron Kohavi and Barry Becker. Adult data set. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996. Dataset available from the UCI Machine Learning Repository.

[7] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Hoboken, 3 edition, 2020. ISBN 978-0-470-44274-6.

[8] Pierre-Emmanuel et al Poulet. Prediction-powered inference for clinical trials. *medRxiv : the preprint server for health sciences 2025.01.15.25320578.*, 2025.

[9] T.E. Raghunathan, J.P. Reiter, and D.B. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16, 2003.

[10] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

[11] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987. ISBN 978-0-471-28700-9.

[12] W. S. Student. The probable error of a mean. *Biometrika*, 6:1–25, 1908.

[13] S. Wang, T. H. McCormick, and J. T. Leek. Post-prediction inference. *[Preprint]. bioRxiv. https://doi.org/10.1101/2020.01.21.914002*, 2020.

[14] Anru Zhang, Lawrence D. Brown, and T. Tony Cai. Semi-supervised inference: General theory and estimation of means. *Annals of Statistics*, 47(5):2538–2566, 2019. doi: 10.1214/18-AOS1756.

[15] Zhongheng Zhang. Missing data imputation: focusing on single imputation. *Annals of translational medicine*, 4(9), 2016.

[16] Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *arXiv preprint arXiv:2309.16598*, 2023.

[17] Tijana Zrnic and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.