Jarren Briscoe, Garrett Kepler,
Daryl Deford, Assefaw Gebremedhin

Washington State
University

Algorithmic Accountability in Small Data

Sample-Size-Induced Bias Within Classification Metrics

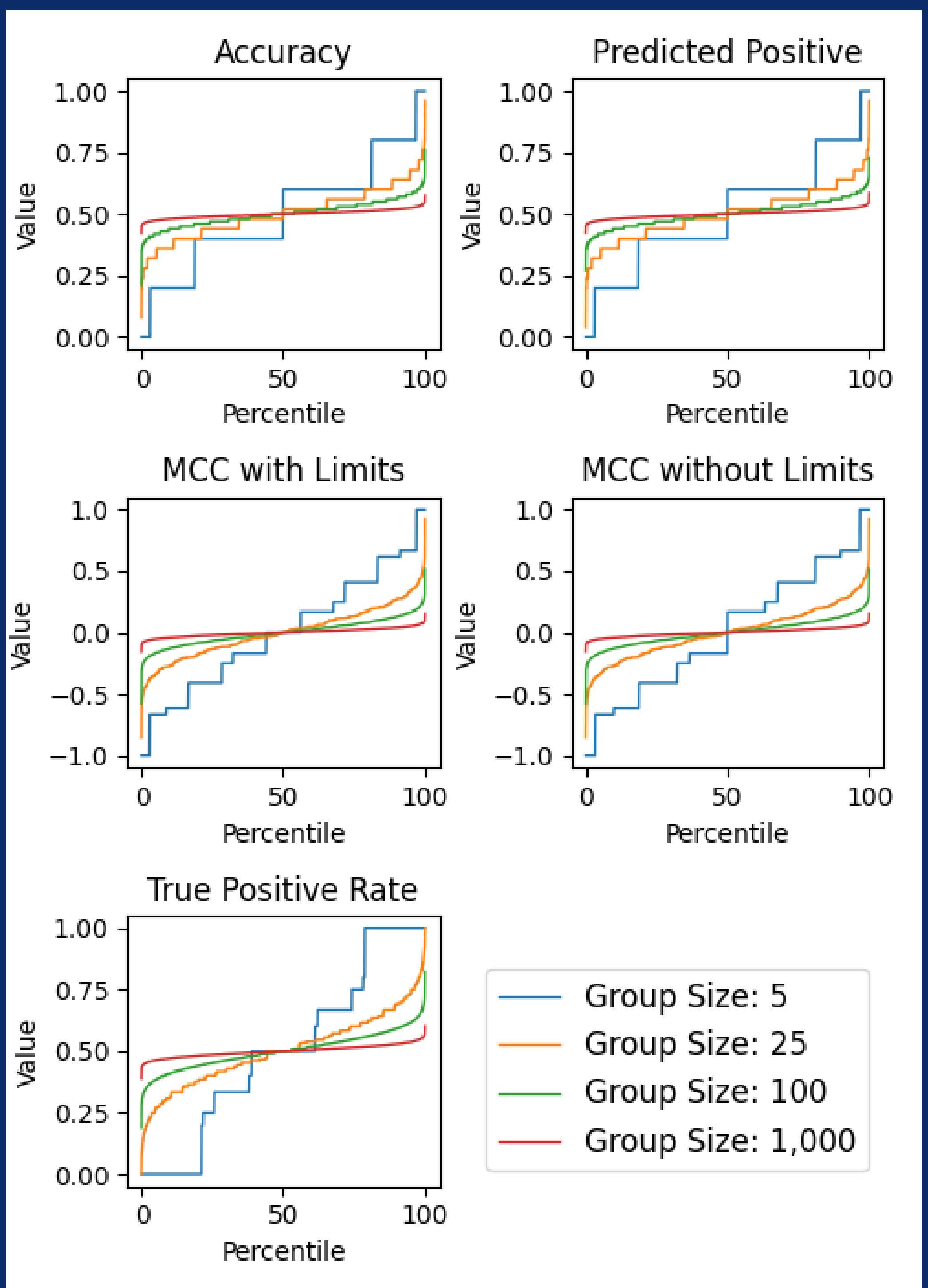
Overview

We rigorously analyze classification metrics on small data. We analyze fifteen classification metrics, including **accuracy, precision,** and **recall**, and discover that these metrics can be **unstable** or **misleading**. We find sample sizes induce **combinatorial bias** and metric variability, with fairness assessment implications. We introduce two methods to address this issue: the MATCH Test, a probabilistic comparison of metric estimates, and Cross Prior Smoothing (CPS), a model-agnostic and data-driven correction technique.

MATCH Test

In fairness audits, group comparisons are often based on scalar metric values without considering the underlying variability. The Metric Alignment Trial for Checking Homogeneity (MATCH) Test addresses this by providing a **probabilistic measure** of whether observed disparities are likely under the same performance assumptions.

Empirical Cumulative Density Functions of Common Metrics



Cross-Prior Smoothing (CPS)

Cross-Prior Smoothing is a model-agnostic technique designed to reduce the volatility of classification metrics caused by small sample sizes. It adjusts a group's confusion matrix by incorporating prior information from a larger, related reference group. This mitigates sharp fluctuations in classification metrics, especially when sample counts are low. CPS works by blending observed counts with scaled-down reference counts using a strength parameter λ, improving both reliability and fairness in comparative analyses—particularly important when one or more demographic subgroups have less than 100 samples.

Experiments

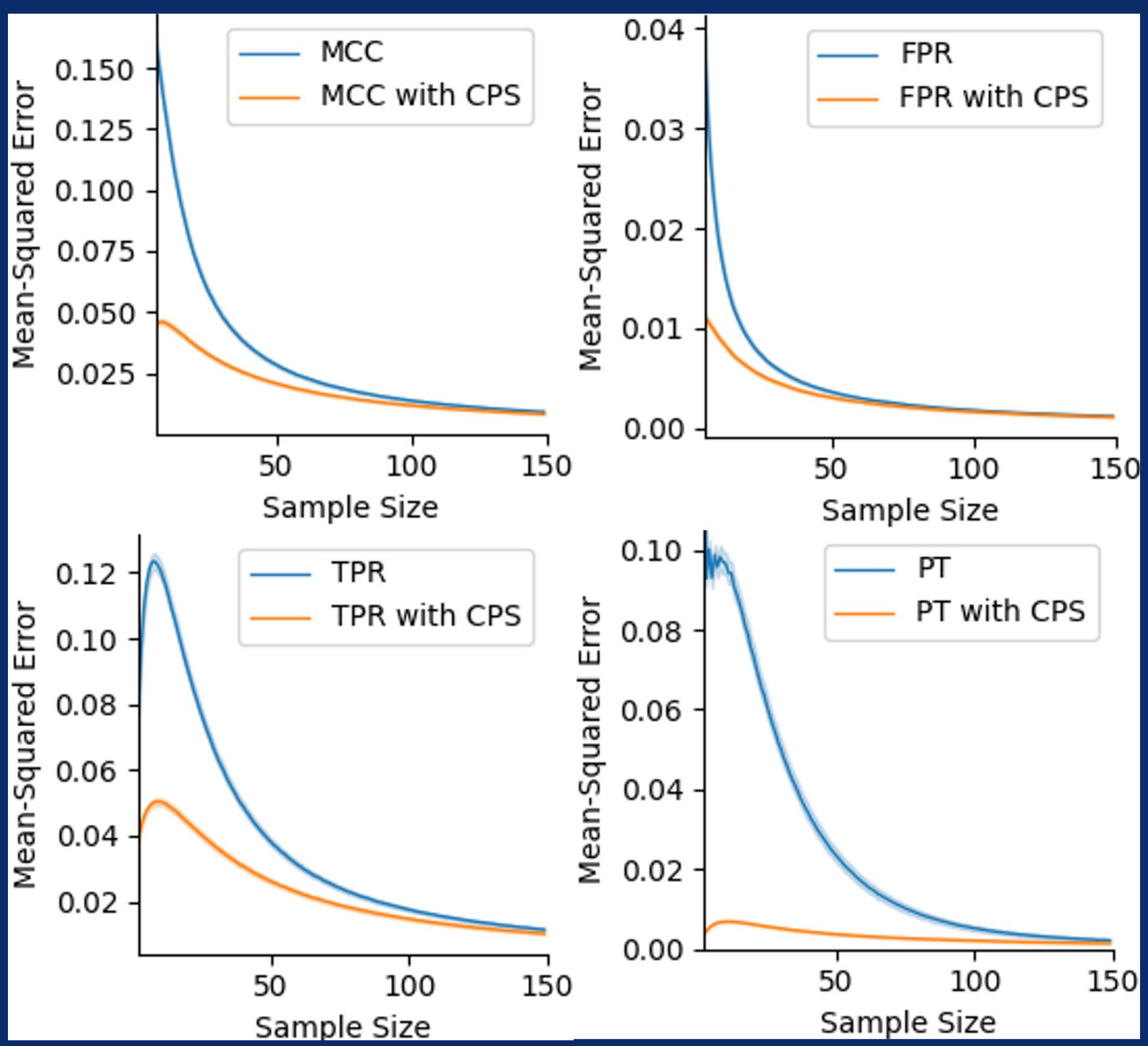
Experiments on COMPAS and Folktables' datasets show that CPS consistently reduces metric error across all 15 tested metrics. Compared to traditional smoothing, CPS consistently yields lower mean-squared error and fewer undefined cases.

Practical Guidelines

1. **Reasonable Prior Assumption:** The reference group is reasonably expected to be informative.
2. **Reference Size:** The reference group should have at least 100 samples.
3. **Choosing λ:** Values in {5, 10, 20} worked well empirically. λ reflects confidence in the reference group, which can be evaluated in many ways.

Sample-Size-Induced Bias Among all Groups:

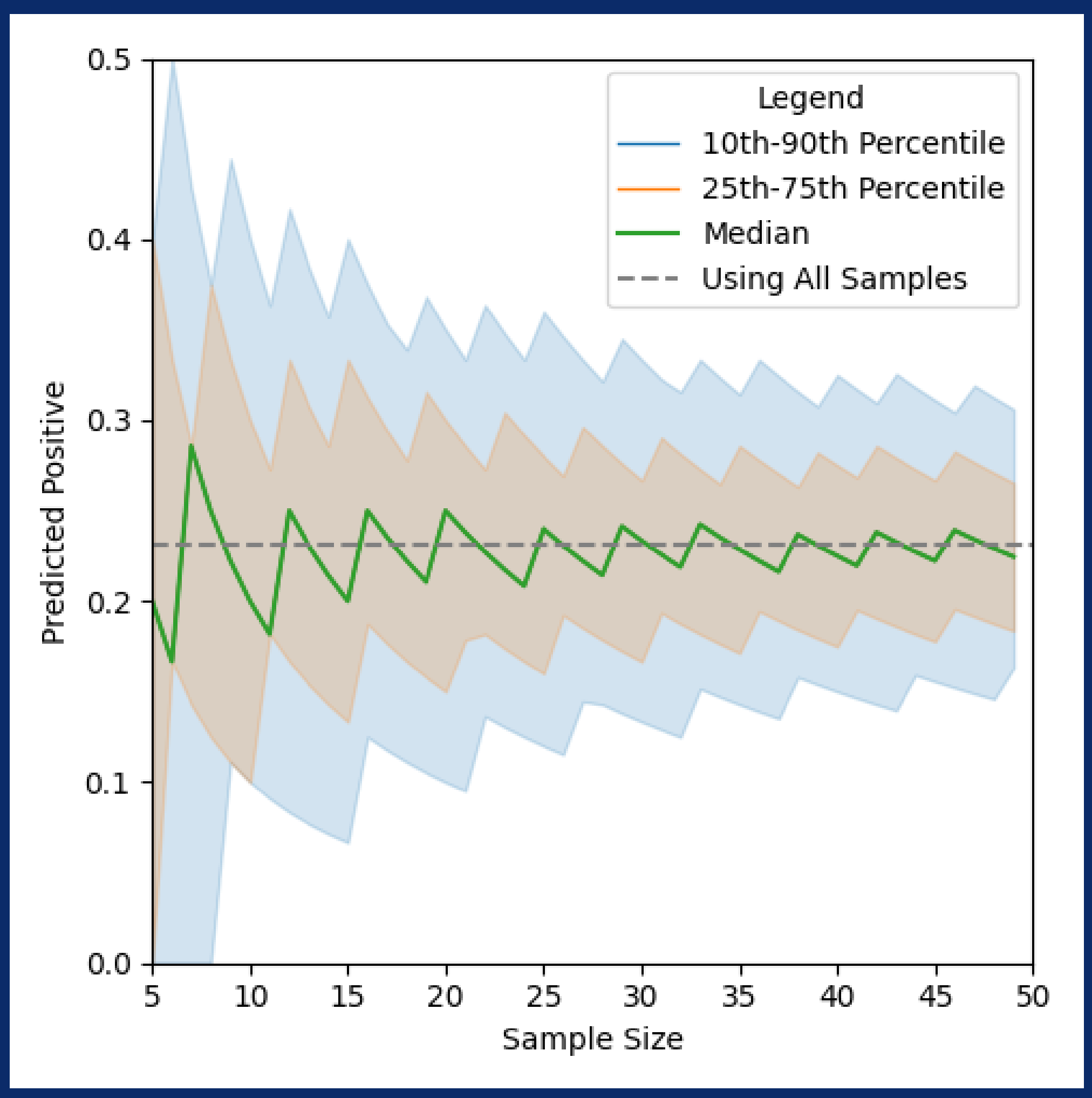
Comparing CPS and Original (With Substitutions for Undefined Cases)



Variability & Jaggedness

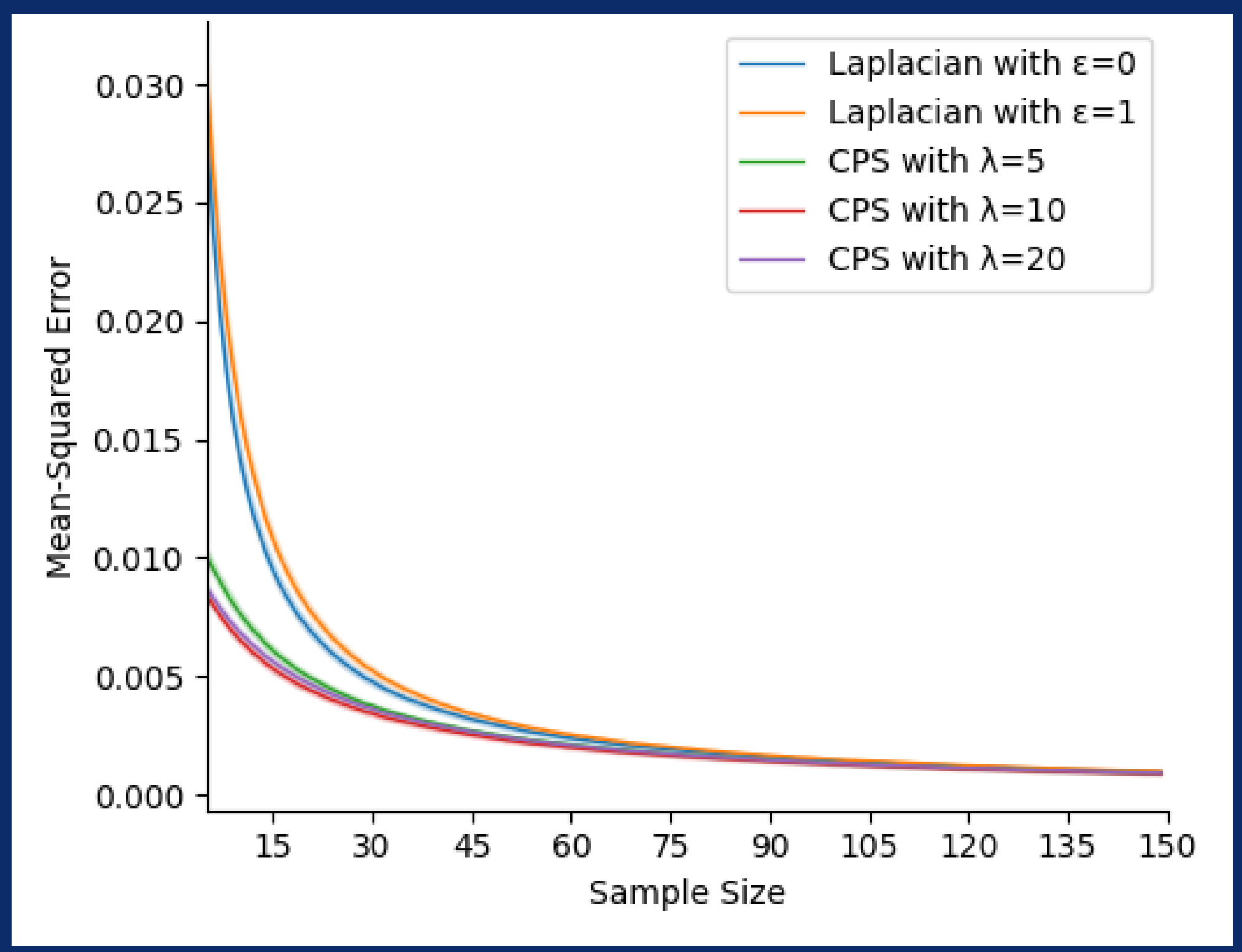
Classification metrics derived from confusion matrices can behave erratically. This phenomenon—**jaggedness**—refers to abrupt, discrete jumps in metric scores (estimates) caused by the combinatorial structure of the matrix space. Even a single observation can significantly shift the estimate distribution. This variability undermines fairness assessments, when comparing groups of unequal size. Our work highlights the theoretical and empirical foundations of this issue and illustrates why naive comparisons across sample sizes can be misleading without accounting for these distributional shifts.

Predicted Positive Rate for Multiracial Individuals in Folktables' Income Dataset



MSE of Prevalence Across Sample Sizes:

Comparing CPS, Laplacian Smoothing, and Original



Undefined Cases

Calculating Counts and Growth Rates

Metric	Asymptotic Growth	Undefined Case Count
Binomial Metrics (ACC, PREV, . . .)	$\Theta(1)$	0
Marginal Benefit	$\Theta(1)$	0
Objective Fairness Index	$\Theta(1)$	0
Simplified F_1 Score	$\Theta(1)$	1
Joint Ratio Metrics (TPR, FPR, . . .)	$\Theta(n)$	$n+1$
Matthews Correlation Coefficient	$\Theta(n)$	$4n$
Prevalence Threshold	$\Omega(n)$ & $\mathcal{O}(n \log\log n)$	$\geq 2n+2$ and $< e^{\gamma} n \log\log n + \frac{0.6483n}{\log\log n}$
Treatment Equality	$\Theta(n^2)$	$\binom{n_1+2}{2} + \binom{n_2+2}{2} - 1$
Original F_1 Score	$\Theta(n^2)$	$\binom{n+2}{2}$
Unique Confusion-Matrices Count	$\Theta(n^3)$	Count of All Possible CMs: $\binom{n+3}{3}$

Conclusions

Classification metrics exhibit jaggedness—erratic jumps caused by the combinatorial structure of confusion matrices. This can distort fairness and performance comparisons across groups.

We propose two solutions:

- **Cross-Prior Smoothing (CPS):** Stabilizes metrics by leveraging an informative prior.
- **MATCH Test:** Assesses whether metric differences are statistically meaningful.

These tools improve metric reliability and support fair evaluations.

Contact Information

jarren.briscoe@wsu.edu
assefaw.gebremedhin@wsu.edu

Acknowledgements

This research is supported by a GAANN fellowship from the U.S. Department of Education.