Addressing rare outcomes in PER quantitative studies

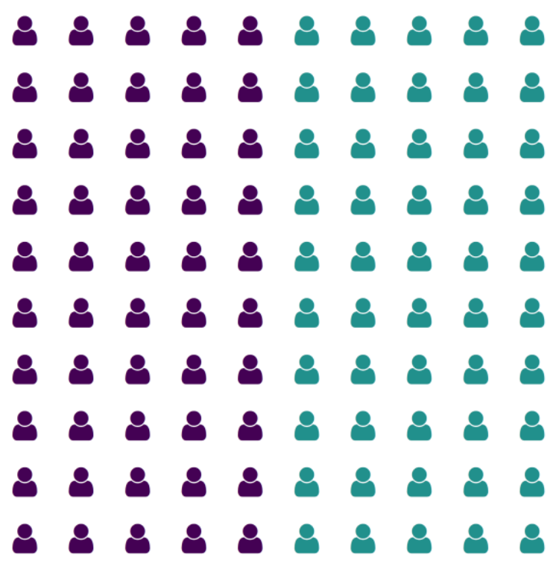
Nicholas T. Young[1,2], Marcos D. Caballero[1,2,3,4]

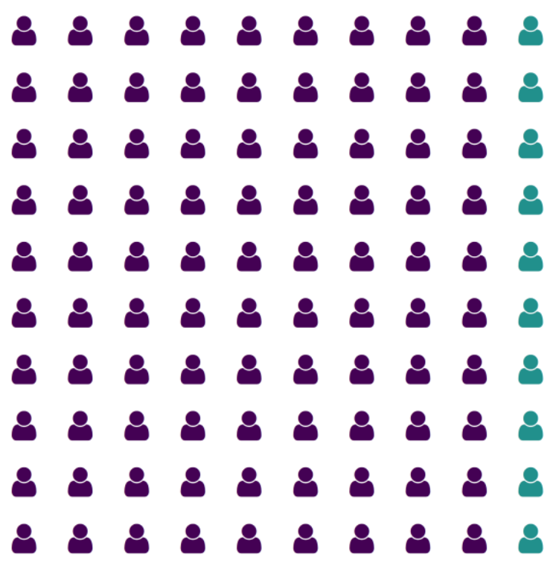
[1]Department of Physics and Astronomy, Michigan State University
[2]Department of Computational Mathematics, Science, and Engineering, Michigan State University
[3]CREATE for STEM Institute, Michigan State University
[4]Department of Physics and Center for Computing in Science Education, University of Oslo

PERL @MSU

CREATE for STEM Institute

Prior work[1] tells us random forest[2,3] will find this feature more predictive than this feature





even if both features are created to be equally predictive!

This is a problem because PER data is more like the right image than the left image.

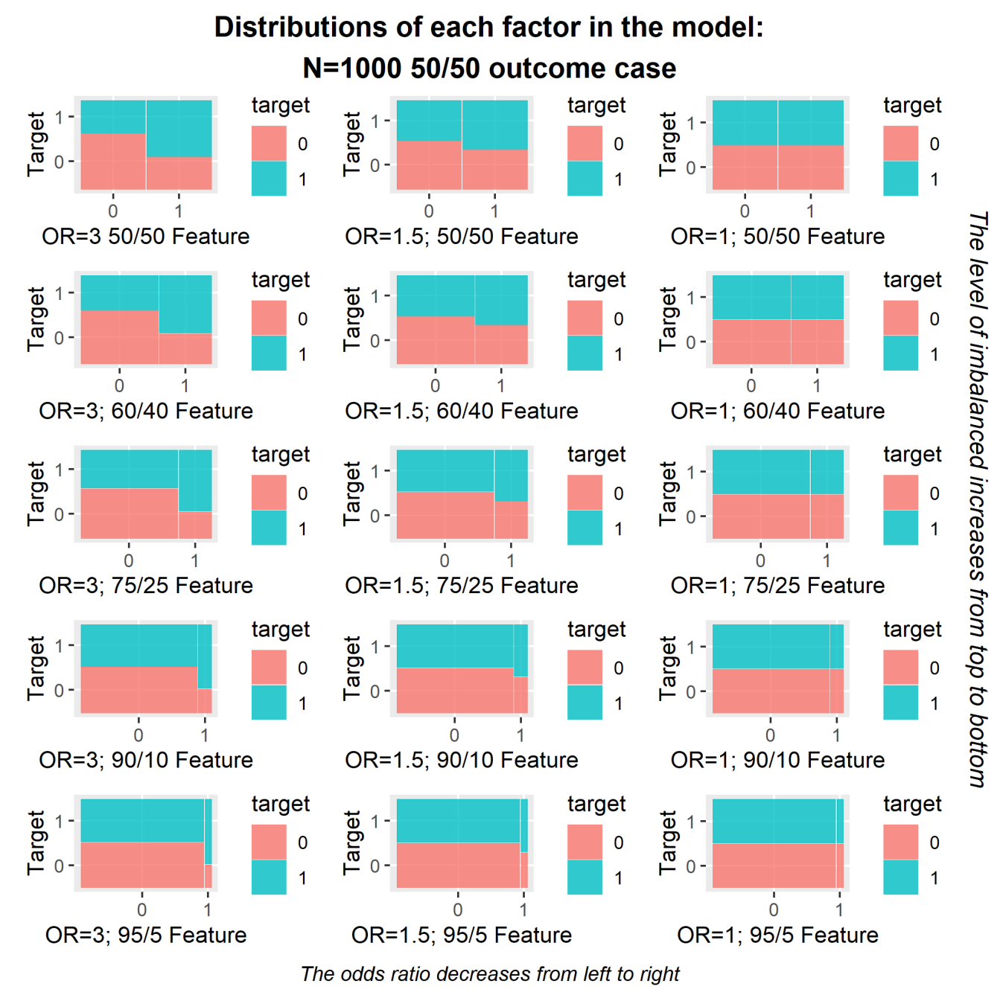
However, that simulation study[1] didn't use

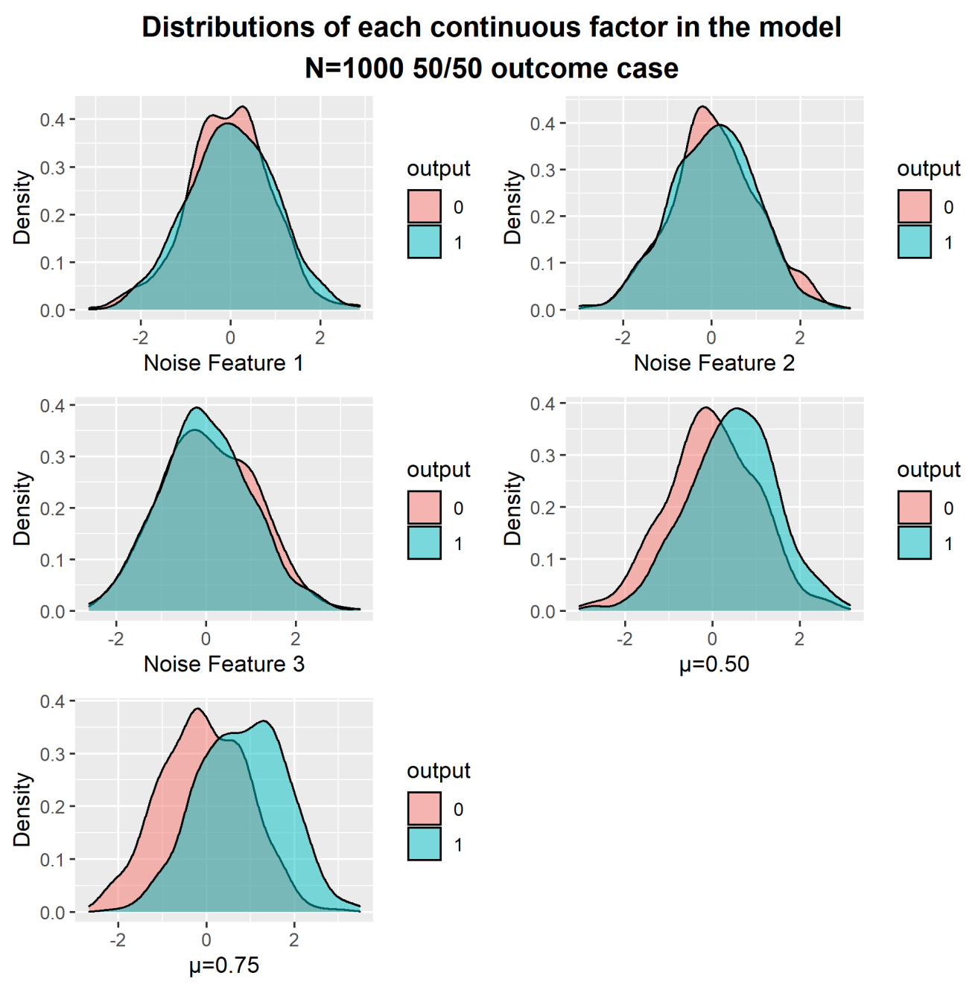
- data combinations typical of PER
- imbalanced outcomes
- newer feature selection approaches[4]

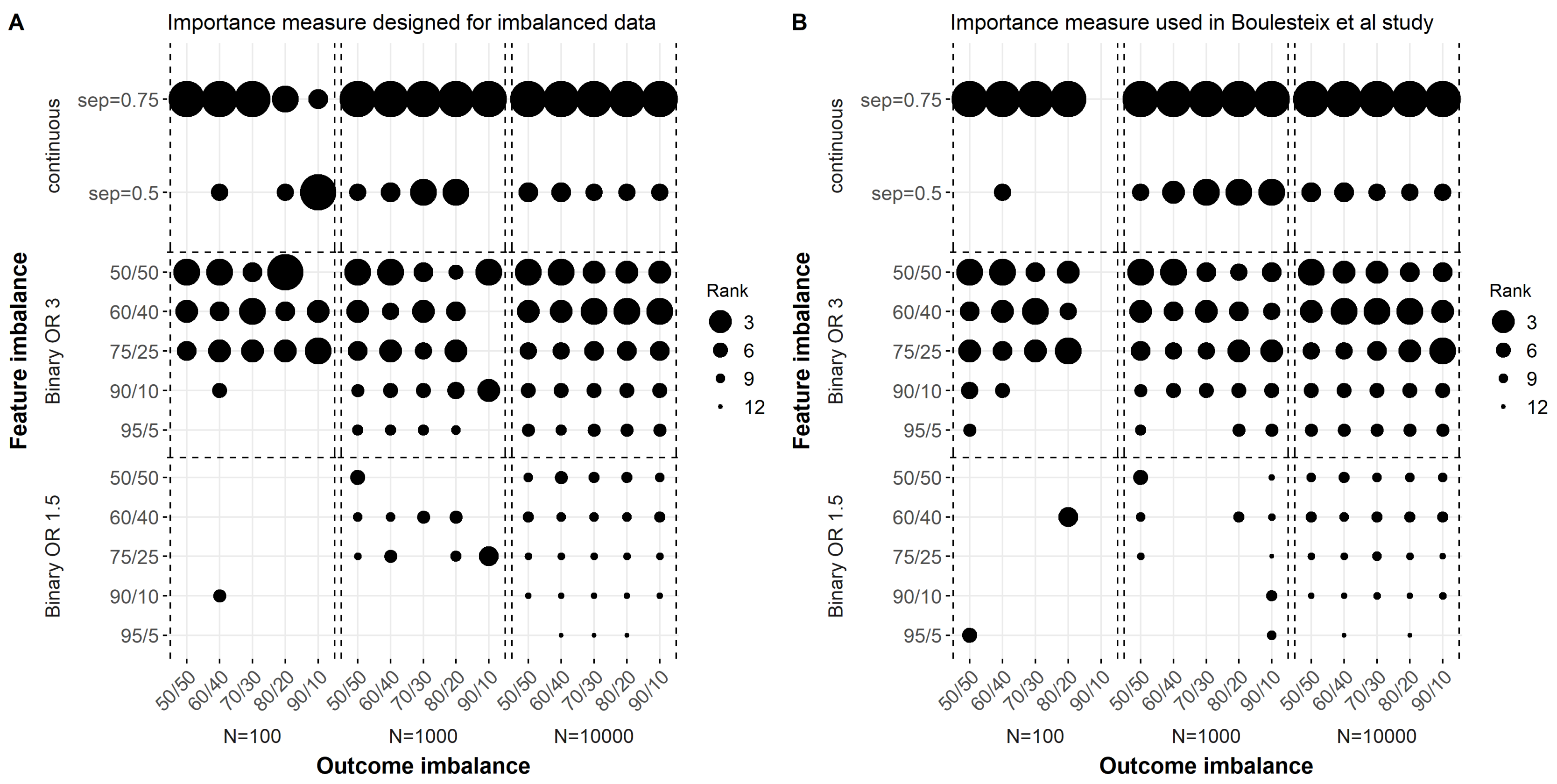
Let's try it ourselves!

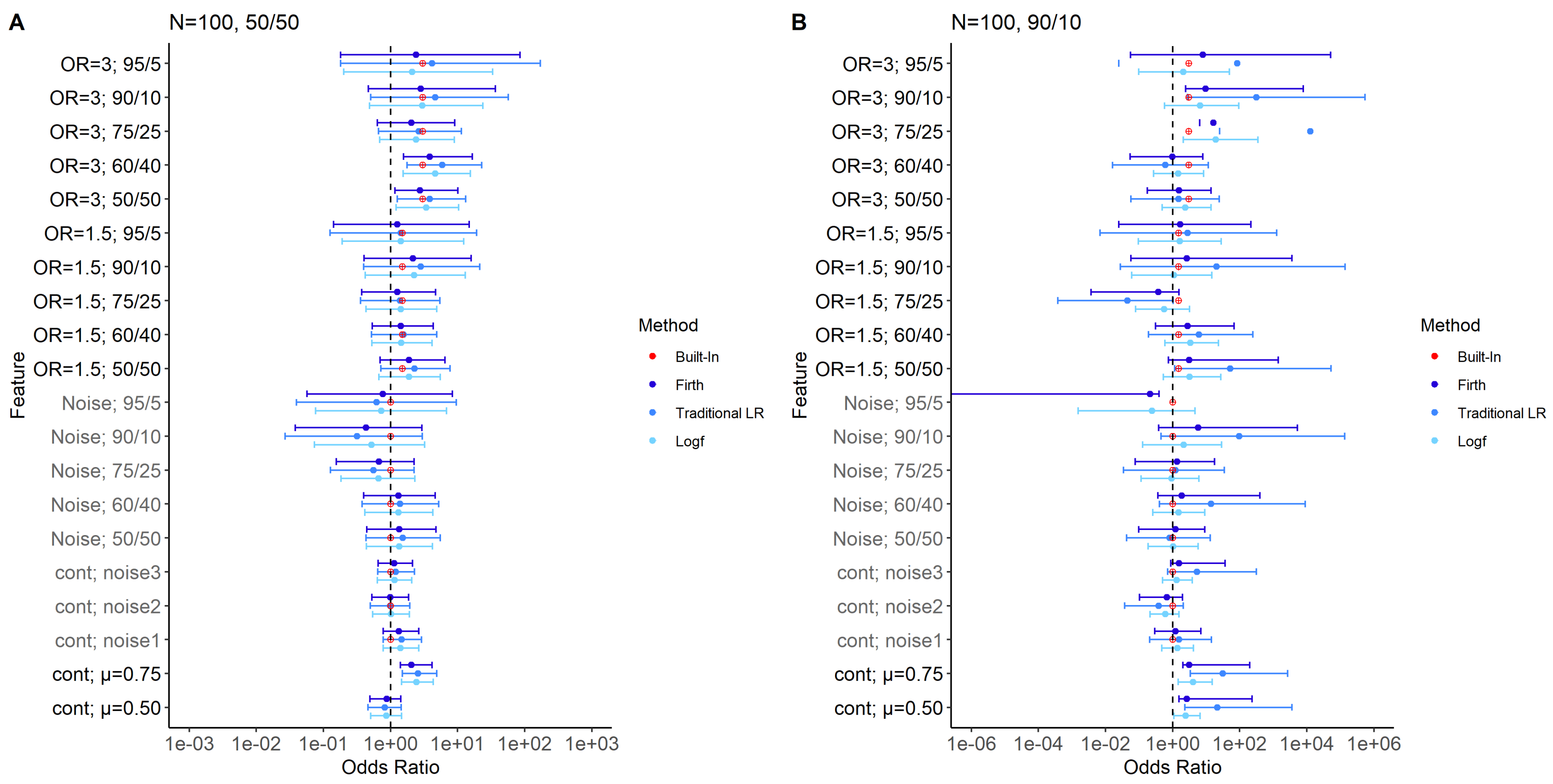
Our data

- *3 levels of predictiveness:* Odds Ratio (OR) = {3, 1.5, 1}
- *5 feature imbalances*: 50/50, 60/40, 75/25, 90/10, 95/5
- *5 continuous features:* 2 informative, 3 noise
- *5 outcome imbalances*: 50/50, 60/40, 70/30, 80/20, 90/10
- *3 sample sizes: N= {100, 1,000, 10,000}*









Lower-imbalance features rank higher than higher-imbalance features for identical OR.

- Results independent of sample size and outcome imbalance
- Feature selection approaches for imbalanced data don't offer an improvement over standard approaches.
- Many predictive features aren't detected for N ≤ 1000.

Logistic regression also has an imbalance bias.

- Confidence intervals can span orders of magnitude for imbalanced features.
- Problem is worse for highly imbalanced outcomes.
- Penalized regression methods such as Firth[5] and Logf[6] can help for smaller samples sizes.

School 1

140 Applicants

41% rejected

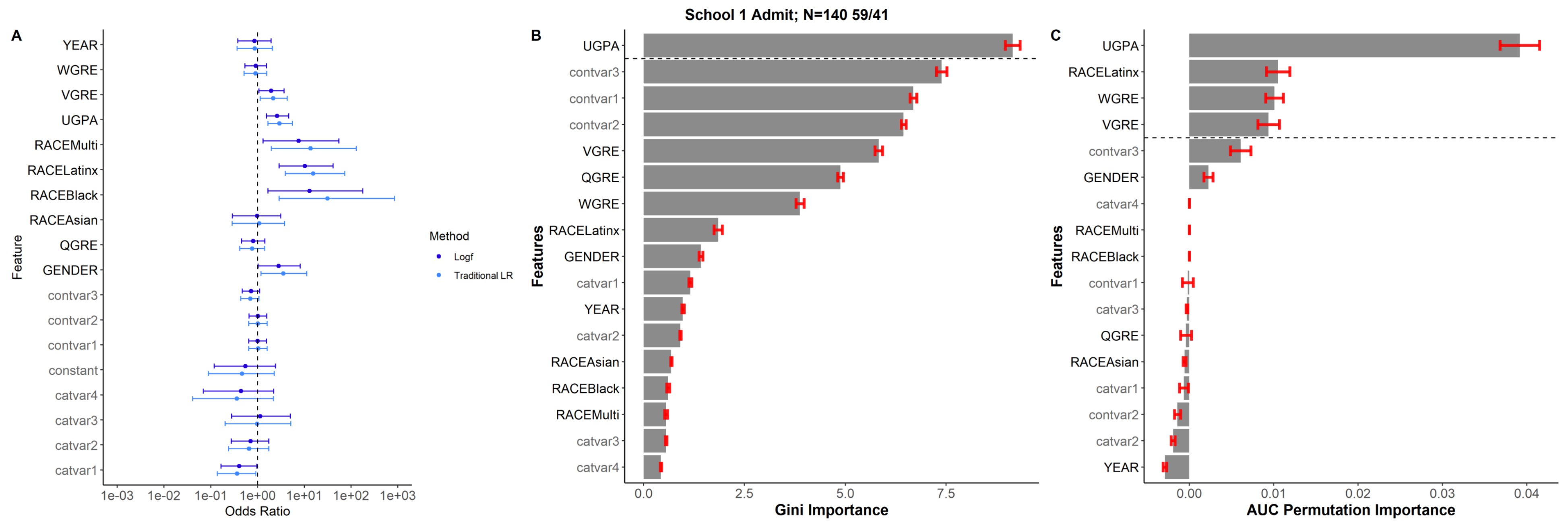
21% women

19% Latinx

13% Asian

4% Black

4% Multiracial



This bias is observed in real PER data.

- Logf considerably shrinks confidence interval for RaceBlack.
- RaceMulti, RaceBlack, RaceLatinx have around the same odds ratio but only RaceLatinx (least imbalanced) is different from noise when using AUC permutation importance.

We might be introducing false negatives into our results due to our data.
Therefore, researchers should report feature and outcome imbalance in their publications.

Funding provided by Michigan State University's College of Natural Science & the Lappan-Phillips Foundation. Data provided by the Inclusive Graduate Education Network.

[1]Boulesteix A.L, Bender A., Bermejo J. L., Strobl C., Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations, Briefings in Bioinformatics, Volume 13, Issue 3, May 2012, Pages 292–304.
[2]Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).
[3]Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. BMC bioinformatics, 8, 25.
[4]Janitza, S., Strobl, C., & Boulesteix, A. L. (2013). An AUC-based permutation variable importance measure for random forests. BMC bioinformatics, 14, 119.
[5]Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. Biometrika, 80(1), 27-38.
[6]Greenland, S., and Mansournia, M. A. (2015) Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. Statist. Med., 34: 3133– 3143.