Evaluating machine learning classifier performance with biostatistical methods from diagnostic testing

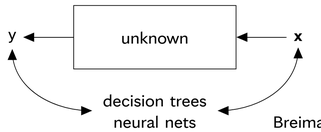
Momin M. Malik, Ph.D.[1], Dave Watson, Ph.D.[2], Madison J. Beenken, M.S.[2], Chung Il Wi, M.D.[1], Young J. Juhn, M.D., M.P.H.[1]

[1] Department of Pediatric and Adolescent Medicine, Mayo Clinic; [2] Department of Quantitative Health Sciences, Mayo Clinic

BACKGROUND

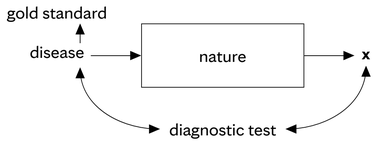
Classification and diagnostic testing

- **Machine learning classification** circumvents causality/understanding/explanation to make reliable input-output mappings [1]. Only care about reliability of mappings.



Breiman's (2001) diagram

- **Diagnostic testing** in medicine [2] is use of a cheap/fast proxy in place of an onerous/expensive gold standard. Only care about how well a diagnostic test works.



- *Methods for evaluating diagnostic tests apply directly to machine learning classification!*
- Apply analytic confidence intervals to metrics in a completely held-out test set [3]
- **Example area: machine learning "fairness"**
  - Software: all **point estimates** of **risk ratios**
    - AI Fairness 360 (IBM)
    - Fairlearn (originally from Microsoft)
    - TensorFlow Fairness Indicators (Google)
    - fairmodels (Warsaw UofTech)
    - Aequitus (Data Science for Social Good)

METHODS

Improve existing approaches

- Use appropriate approximate intervals; correct for multiple comparisons
- Do better "fairness audits" with analytic confidence intervals for unpaired risk ratios
  - Koopman asymptotic score confidence intervals have the best coverage [4,5]

Introduce better approaches

- Non-significant results may point to an underpowered study, rather than "fairness"
- Pick a small set of substantively motivated comparisons and metrics, especially for interactions, rather than lose power
- Use other appropriate analytic approaches, e.g., Baptista-Pike mid-*p* asymptotic score intervals for odds ratios of unpaired data (instead of logit intervals or bootstrapping)

Extensions (not shown)

- Clustered data: Generalized estimating to get appropriate odds ratios CIs
- Non-inferiority: show between groups, or, use for model selection (e.g., significantly less disparate performance for no significant difference in performance)

Application

- Re-apply to point estimates in our study [6] on bias by SES in a model of childhood asthma, and with better chosen comparisons

RESULTS

Improve current approaches

Uncertainty quantification shows nothing is significant, in contrast to the original work [6]

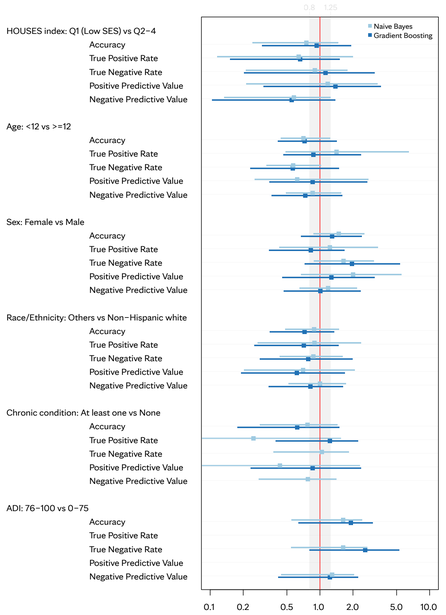


Figure 1. Risk ratios (log scale), with Bonferroni-corrected Koopman asymptotic score confidence intervals.

Introduce better approaches

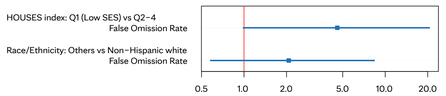


Figure 2. Odds ratios (log scale), with Baptista-Pike mid-p interval asymptotic score confidence intervals for a substantive metric [7] and groups is better but still not significant

CONCLUSIONS

- Technical aspects of model auditing are superseded by concerns in system purpose and goals, power imbalances, audit study designs, actionability, and accountability; but existing methods can be applied to easily improve technical aspects of "fairness"
- Further lessons and methods to import: "index test" limits the sampling frame; case-control data (sampling on the dependent variable) is usable but in limited ways

This work was funded by NIH R01 HL171508, Mayo Clinic Health System Research, and the Mayo Clinic Children's AI Program.

REFERENCES

1. Breiman L. Statistical modeling: the two cultures. Stat Sci. 2001 Aug;16(3):199-231.
2. Zhou X-H, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. John Wiley & Sons, Inc.; 2011.
3. Bates S, Hastie T, Tibshirani R. Cross-validation: what does it estimate and how well does it do it? J Am Stat Assoc. 2023 May;119(546):1434-1445.
4. Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. Stat Methods Med Res. 2015 Apr;24(2):224-54.
5. Fagerland MW, Lydersen S, Laake P. Statistical analysis of contingency tables. Chapman and Hall/CRC; 2017.
6. Juhn YJ, Ryu E, Wi CI, King KS, Malik M, Romero-Brufau S, et al. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. J Am Med Inform Assoc. 2022 Jun 14;29(7):1142-1151.
7. Rodolfa K, Saleiro P, Ghani R. Bias and fairness. In: Foster I, Ghani R, Jarmin RS, Kreuter F, Lane J, editors. Big data and social science. 2nd ed. Chapman and Hall/CRC; 2021. p. 281-312.