Statistically Valid Machine Learning Fairness Evaluation

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

ABSTRACT

BACKGROUND

Within the machine learning "fairness" literature, models 'audits' are done via point estimates (ratios of metrics like sensitivity between subgroups). But:

- Estimates of model performance or fairness are really estimators of possibly out-of-sample performance.
- For a given sample size, an effect size may be consistent with what we would expect from random variation.
- As estimators, model metrics have theoretical distributions from which we can determine statistical significance, as well as look at how these estimators may become statistically biased by certain study designs or features of data.

OBJECTIVE

The well-developed frameworks in biostatistics for diagnostic testing, and methods in statistics for contingency tables, apply directly to AI/ML classifiers more generally; these principles can be extended to do statistically valid fairness evaluation.

METHODS

We determined what are relevant statistical and biostatistical topics, and made original linkages to machine learning methodology. From this, we produced set of recommended analytic statistical methods and study designs for doing fairness testing.

RESULTS

- If using risk ratios (the standard in machine learning, and what is produced in all the major ML fairness software packages), Koopman asymptotic score confidence intervals are a good choice.
  - However, odds ratios have better properties across the range of estimated probabilities, and may be preferred.
  - If our test statistics are odds ratios, we can do performance comparisons for clustered data via Generalized Estimating Equations.
- If making multiple comparisons (across subgroups, models, before-and-after for model corrections), a correction should be applied (e.g., Bonferroni, False Discovery Rate, etc.).
- A "data mining" approach to fairness will rapidly deplete statistical power, especially if also testing multiple metrics at intersections of group membership (e.g., race and gender). It is better practice to substantively choose the relevant metric(s), and relevant subgroups to compare.

CONCLUSIONS

Closing the gap between practices that come out of machine learning and well-understood statistical methods and approaches will result in more reliable and meaningful work.

OBJECTIVES

- Major fairness auditing software packages all use the (arguably inappropriate[1]) "four-fifths" rule for determining the relevance of ratios of machine learning model performance metrics .
  - Includes IBM's AI Fairness 360, Fairlearn (originally from Microsoft), Google's TensorFlow Fairness Indicators, and independent tools like fairmodels and Aequitus.
  - None quantify uncertainty or do testing
- Technical aspects of "fairness auditing" are superseded by concerns in system purpose and goals, power imbalances, audit study designs, actionability, and accountability;[2] but improving technical aspects is straightforward.

METHODS

Existing analytic work covers uncertainty quantification, and aspects of study design. Relevant areas are statistical methods for diagnostic testing; the statistical analysis of contingency tables;[3] statistical methods for ROC curve analysis; and machine learning work on cross-validation.[4]

RESULTS

Data splitting

- *k*-fold cross validation is unreliable for estimating out-of-sample performance; better is to use a completely held-out test set, that is ideally used only once.[4] However, *k*-fold cross validation is consistent for model selection. Uncertainty need not be calculated as a part of model selection, but should be calculated for final estimates of model performance or fairness on the completely held-out test set. Data splitting should also be done to have dependent data within the same fold of data. E.g.,
  - all records associated with a given patient should be within the same fold;
  - if there is variability in practice site/office, all patients for the same practice should be in the same fold;
  - for temporal data, folds should be strictly orderable in time; etc.

Clustered data

- Generalized Estimating Equations are a closed-form way of estimating model performance (e.g., multiple observations per patient). However, they can only give odds ratios, not risk ratios.

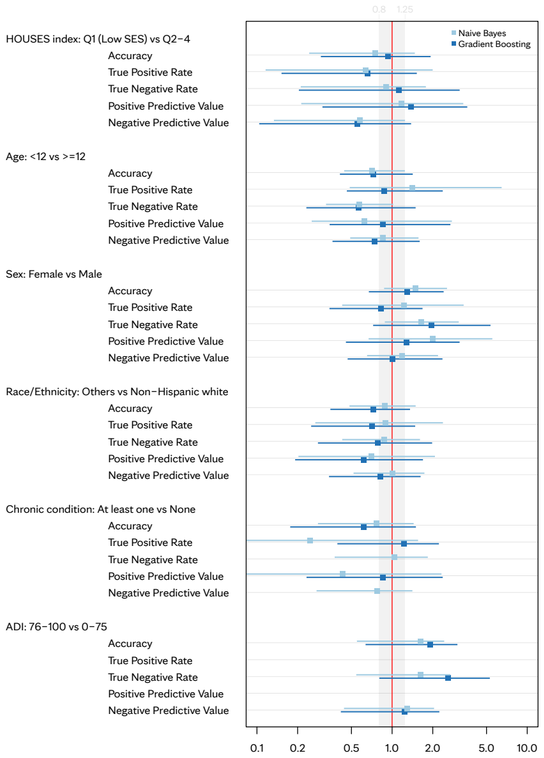
Risk ratios

- Estimated probabilities are also known as binomial proportions, with ratios called *risk ratios* or *relative risk* (although this name is not known or used in machine learning literature). Risk ratio confidence intervals can be used for machine learning fairness.
- Like with the Agresti-Coull interval for binomial proportions, approximate intervals interpolate over the discreteness of both count data and of small sample size and are better than exact intervals (and bootstrapping). Across proposed intervals, Fagerland et al.[3] find that the 1984 Koopman asymptotic score interval has the best coverage (95% intervals contain the true parameter at least 95% of the time). This interval is complicated to calculate, but implementations exist in statistical software.

RESULTS, cont.

Re-doing, with proper uncertainty quantification, the results of our work[5] examining socioeconomic bias in a model for pediatric asthma exacerbation, we find nothing was significant (Fig. 1).

FIGURE 1: Risk Ratios (log scale), with Bonferroni-adjusted Koopman asymptotic score CIs.

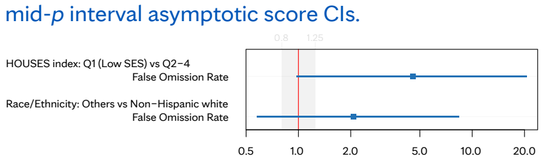


While our original paper[5] used the common "four-fifths rule" of risk ratios less than 0.8 to determine fairness or unfairness[1] across 60 sets of subgroup comparisons by various metrics, properly quantifying the uncertainty around these shows that none of the metrics are statistically significant.

Recommended: Odds ratios, and limited comparisons

- As risk ratios only apply to small probabilities, odds ratios are preferable. Baptista-Pike mid-*p* asymptotic score confidence intervals work well.[3]
- Substantively choosing a small set of comparisons obviates the need for correction (Fig. 2). Loss of statistical power is worse if we also want to test for intersections of subgroups, making it even more important to avoid uninformative comparisons.

FIGURE 2: Odds ratios (log scale) with Baptista-Pike mid-*p* interval asymptotic score CIs.



A smaller set of comparisons, using a substantively chosen metric (false omission rate, capturing who is left out of a potentially assistive intervention[6]) and substantively chosen subgroup comparisons, does not require a correction for multiple comparisons that will decrease statistical power. Still, insufficient sample size will result in non-significant results.

DISCUSSION

Using the re-analysis of our own previously published work[5] as an example, we show the importance of quantifying uncertainty. We both used an appropriate method for risk ratios, and also applied a Bonferroni correction for the multiple comparisons. When results are not significant, the conclusion should not necessarily be that the model is actually "fair", but potentially that the sample size is insufficient to make a conclusion one way or the other. This is the case for our study, where had relatively small sample size.

For future studies that plan to gather data by which to evaluate machine learning fairness, the formulae for confidence intervals can be used to do sample size and power calculations. This will determine what are needed sample sizes (especially for marginalized groups, and for intersections of group memberships) to carry out meaningful fairness testing.

CONCLUSIONS

Well-understood and rigorously tested statistical frameworks exist for evaluating models that estimate probabilities, such as classification models, or for some negative outcome. Rather than ignoring century-old lessons about the importance of uncertainty quantification, or using computationally inefficient and poorly performing approaches like bootstrapping, adopting existing analytic methods and study design guidelines (including doing sample size calculations) will result in more reliable and meaningful work.

ACKNOWLEDGMENTS

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

REFERENCES

1. Watkins EA, Chen J. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. Proc ACM Fairness Account Transparency (FAccT); 2024, 764-775. doi: 10.1145/3630106.3658938.
2. Raji ID. The anatomy of AI audits: form, process, and consequences. In: Bullock JB, Chen Y-C, Himmelreich J, Hudson VM, Korinek A, Young MM, Zhang B. The Oxford handbook of AI governance. Oxford: Oxford University Press; 2023, 495-516. doi: 10.1093/oxfordhb/9780197579329.013.28.
3. Fagerland MW, Lydersen S, Laake P. Statistical analysis of contingency tables. Chapman and Hall/CRC; 2017. doi: 10.1201/9781315374116.
4. 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. doi: 10.1080/01621459.2023.2197686.
5. 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. doi: 10.1093/jamia/ocac052.
6. 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. doi: 10.1201/9780429324383-11.