NeurIPS Version

Journal Version

Human Expertise in Algorithmic Prediction

Rohan Alur, Loren Laine, Darrick K. Li, Dennis Shung, Manish Raghavan & Devavrat Shah

MIT EECS — Electrical Engineering | Computer Science | Artificial Intelligence + Decision-making

Yale SCHOOL OF MEDICINE

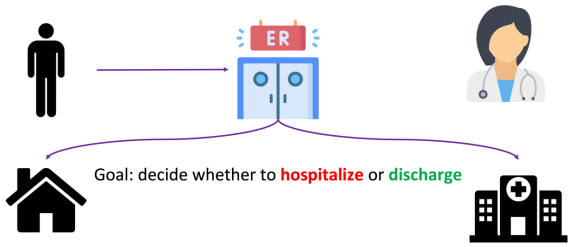
Long known that **algorithmic predictions** usually **outperform human experts**...

"There is no controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one." (Meehl, 1986)

...but **human discretion** still plays a large role in most high-stakes predictions (e.g., clinical triage)

Why?

Case Study: Emergency Room Triage

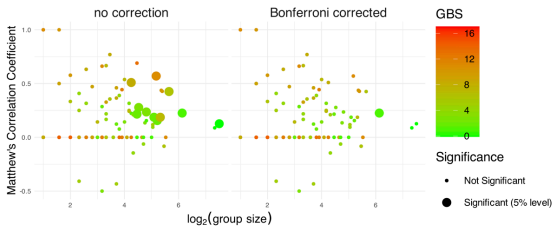


We observe

- $X \in \mathbb{R}^d$ (patient characteristics)
- $\hat{Y} \in \{0, 1\}$ (physician decision)
- $Y \in \{0, 1\}$ (adverse outcome)

But physician has **significantly more information!**

- E.g., can directly examine the patient



Each point represents a group of patients whose features are *identical*. Within each group, we plot the correlation between physician decisions and adverse outcomes. This provides suggestive evidence that physicians incorporate information not encoded in X, but tiny sample sizes preclude meaningful inference.

Do experts incorporate **information** that is **unavailable** to any predictive algorithm?

If so, how can we **leverage human expertise** in prediction tasks?

Algorithmic Indistinguishability

Idea: given class of predictors $\mathcal{F}$, partition inputs such that no $f \in \mathcal{F}$ can **distinguish** between positive and negative instances

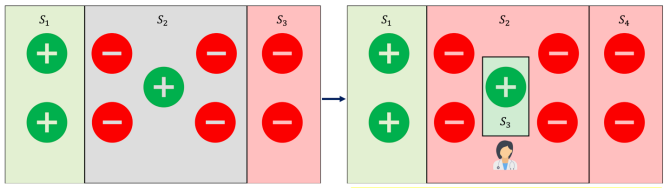
Definition: $S \subseteq \mathcal{X}$ is α-indistinguishable for $\alpha \geq 0$ if

$$|\mathrm{Cov}(Y, f(X) \mid X \in S)| \leq \alpha \; \forall f \in \mathcal{F}$$

Interpretation: no predictor $f \in \mathcal{F}$ is informative within $S \subseteq \mathcal{X}$

$\Rightarrow$ This is essentially *multicalibration* (Hébert-Johnson et al., 2017; Gopalan et al., 2021)

Ideally, expert can provide **additional signal** within **S**



Example: no hyperplane can distinguish positive/negative instances within each subset

Maybe doctor can distinguish patients who "look the same" to algorithmic predictors?

Key Result

Simple predictors (e.g., linearly regressing Y on $\hat{Y}$) **provably outperform** *any* $f \in \mathcal{F}$ *within* each indistinguishable subset

Even if $\mathcal{F}$ contains complex, nonlinear predictors
Even if $\hat{Y}$ is *less* accurate than the best predictor $f^* \in \mathcal{F}$

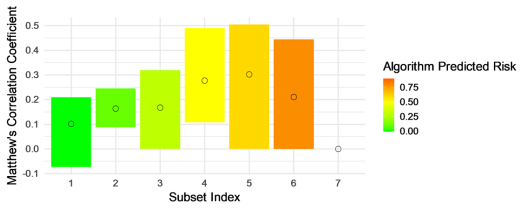
Suggests simple method for incorporating expertise

1. Commit to a model class $\mathcal{F}$
2. Find indistinguishable subsets $S_1, \ldots, S_K \subseteq \mathcal{X}$
3. Use $\hat{Y}$ to predict Y within each subset

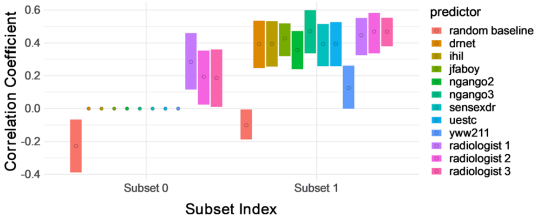
See paper for details and extensions

- How do we identify indistinguishable subsets?
- Real-valued outcomes, vector-valued human feedback (e.g., clinical notes)
- Information-theoretic interpretation of results
- Predictors which are robust to human noncompliance

Experiments



Physician triage performance within subsets which are indistinguishable with respect to class of depth ≤ 3 regression trees. For two subsets, representing ~24% of patients, physician judgment provides signal that these algorithms cannot replicate



Physician diagnostic performance within subsets which are indistinguishable with respect to 8 leaderboard algorithms for diagnosing atelectasis. For one subset, representing ~30% of patients, radiologist judgment provides signal that the algorithms cannot replicate