Score Design for Multi-Criteria Incentivization

Anmol Kabra (anmol@ttic.edu) [1] Mina Karzand [2] Tosca Lechner [3] Nathan Srebro [1] Serena L. Wang [4]

[1]TTI-Chicago [2]UC Davis [3]UWaterloo [4]UC Berkeley

Hospital rating systems

Many agencies rate hospitals based on performance metrics, e.g. disease mortality rates, readmission rates. Agencies can have different goals:

- **Medicare Star Rating.** Rate as many hospitals as possible on a simple, transparent framework.
- **Leapfrog Safety Grade.** High grade if meets minimum safety standards.
- **U.S. News Rankings.** Rankings based on specializations' breadth and depth.

Example rating methodology. Medicare compiles 100+ metrics for each hospital, picks ≈50 metrics with high correlation to all, and computes a weighted average to rate on a 5-star scale.

Goodhart's law in rating systems

Rating systems are prone to strategic behavior by the entities being rated. Goodhart's law (a.k.a. Campbell's law, Cobra effect) says that

"When a measure becomes a target, it ceases to be a good measure."

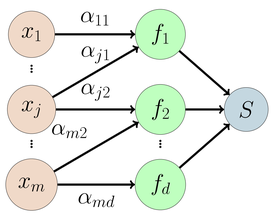
Koretz (2017) notes this strategic behavior in healthcare:

"... a US healthcare provider, PacifiCare of California, set up a system that included incentives for five practices: screening for cervical and breast cancer, checking hemoglobin and A1C for diabetics, and childhood immunizations. Just like a test: the measured practices (the test items) would serve as a sample of all important aspects of practice (the domain). ... Some of the unrewarded practices were unchanged, but others, such as appropriate use of antibiotics and screening for chlamydia, actually deteriorated."

Designing score function given strategic behavior

Kleinberg and Raghavan (2020) give a model for score functions that influence improvement under strategic behavior.

- Hospital invests effort in m activies upto budget B, i.e. $\sum_{j=1}^{m} x_j \le B$.
- Rating agency observes hospital's features $f_1, \dots, f_d$.
- Feature f_i is influenced by effort as $f_i = g_i\left(\sum_{j=1}^{m} \alpha_j x_j\right)$, where g_i is a known, monotonic function.



Rating agency gives a 1-dimensional score function $S(f_1, \dots, f_d)$ and wants to encourage a specific *effort profile* $x^* = (x_1^*, \dots, x_m^*)$.

Issues with the model

1. 1-d score function. But agencies use multiple scores to incentivize.
2. Considers effort x^* that is optimal for score function. Consequences of improving on all scores?

Multi-dimensional score function

We consider a multi-dimensional generalization of the model by Kleinberg and Raghavan (2020).

Performance metrics. d metrics arranged as vector $\boldsymbol{f} \in \mathbb{R}^d$. Agency gets the set of possible performance metrics $\mathcal{F} \subseteq \mathbb{R}^d$.

Score function. Agency designs $S : \mathcal{F} \to \mathbb{R}^k$, and hospital is incentivized to improve and optimize scores $S(\boldsymbol{f})$.

#performance metrics d is large. Want concise score function, i.e. small k.

Score design objectives

Rating agency could have two kinds of objectives for the score function: (1) **Improvement** and (2) **Optimality**. Improvement goal is motivated by agency's safety considerations, and Optimization goal is motivated by hospital's best-response strategy.

1. **Improvement.** Improving on scores $\implies$ Improving on metrics.
$$\text{For all } \boldsymbol{f}, \boldsymbol{f}' \in \mathcal{F}, \text{ if } S(\boldsymbol{f}') \ge S(\boldsymbol{f}) \text{ then } \boldsymbol{f}' \ge \boldsymbol{f}$$

2. **Optimality.** Pareto-optimal scores $\implies$ Pareto-optimal metrics.
$$\text{ParetoOpt}(S) \subseteq \text{ParetoOpt}(\mathcal{F})$$

Note: $\text{ParetoOpt}(S)$ is the set of all $\boldsymbol{f} \in \mathcal{F}$ s.t. no other $\boldsymbol{f}'$ dominates $\boldsymbol{f}$ in all coordinates and strictly-dominates in 1+ coordinates.

Score design restrictions

Agency wants to create S that are "simple and interpretable". We call such requirements "design restrictions".

1. **Coordinate Selection.** The k scores are taken from the d metrics.
$$\text{for all } i \in [k], \text{ there exists } j \in [d] \text{ s.t. } S(\boldsymbol{f})_i = f_j \text{ for all } \boldsymbol{f} \in \mathcal{F}$$
$$\text{or } S : \boldsymbol{f} \mapsto \boldsymbol{A}\boldsymbol{f} \text{ s.t. matrix } \boldsymbol{A} \in \mathbb{R}^{k \times d} \text{ has 1-hot rows}$$

2. **Linear and Monotone.** Improving on the d metrics improves k scores.
$$S(\boldsymbol{f}) = \boldsymbol{A}\boldsymbol{f} \text{ for } \boldsymbol{A} \in \mathbb{R}^{k \times d} \text{ s.t. for all } \boldsymbol{f}, \boldsymbol{f}' \in \mathcal{F},\ \boldsymbol{f}' \ge \boldsymbol{f} \implies S(\boldsymbol{f}') \ge S(\boldsymbol{f})$$

3. **Linear.** k scores are a linear combination of the d metrics.
$$S(\boldsymbol{f}) = \boldsymbol{A}\boldsymbol{f} \text{ for } \boldsymbol{A} \in \mathbb{R}^{k \times d}$$

Task for the agency: Minimal design problem

Given $\mathcal{F} \subseteq \mathbb{R}^d$ × design objective × design restriction, the agency asks:

How to design $S : \mathcal{F} \to \mathbb{R}^k$ to satisfy objective and restriction?

How small can k be?

Design for Improvement goal

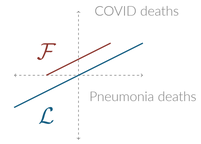
Theorem. Let columns of $\boldsymbol{Z}$ be an orthonormal basis of linear subspace $\mathcal{L}$ associated with $\text{aff}(\mathcal{F})$. For each design restriction, there exists $S : \mathcal{F} \to \mathbb{R}^k$ that satisfies the improvement objective with the following dimensionalities.

Res.	Dimensionality $k \ge$
CS	$\text{ConeSubsetRank}(\boldsymbol{Z}) := \min_q \{q \mid \mathcal{K}_Z = \mathcal{K}_V \text{ for } \boldsymbol{V} \in \mathbb{R}^{q \times r} \text{ s.t. } \boldsymbol{V} \subseteq \boldsymbol{Z}\}$
LM	$\text{ConeGeneratingRank}(\boldsymbol{Z}) := \min_q \{q \mid \mathcal{K}_Z = \mathcal{K}_V \text{ for } \boldsymbol{V} \in \mathbb{R}^{q \times r}\}$
L	$\text{ConeRank}(\boldsymbol{Z}) := \min_q \{q \mid \mathcal{K}_Z \subseteq \mathcal{K}_V \text{ for } \boldsymbol{V} \in \mathbb{R}^{q \times r}\}$

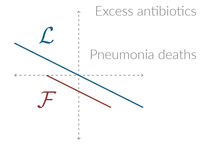
Assume metrics $\mathcal{F} \subseteq \mathbb{R}^d$ have non-empty relative interior with respect to $\text{aff}(\mathcal{F})$. Then the listed dimensionalities k are necessary.

Design algorithm:

1: Given: performance metrics $\mathcal{F}$ and a design restriction.
2: Find $\boldsymbol{Z}$ whose columns are an orthonormal basis of subspace $\mathcal{L}$ associated with $\text{aff}(\mathcal{F})$.
3: Find $\boldsymbol{V}$ that attains the matrix rank corresponding to the design restriction.
4: Find $\boldsymbol{A}$ that satisfies $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{Z}$ and design $S : \boldsymbol{f} \mapsto \boldsymbol{A}\boldsymbol{f}$.



(a) When the two metrics are correlated, we can choose either metric in $S : \mathcal{F} \to \mathbb{R}^1$.



(b) When the two metrics are anti-correlated, we must choose both metrics in $S : \mathcal{F} \to \mathbb{R}^2$.

Figure 2. To design scores for two metrics ($\mathcal{F} \subseteq \mathbb{R}^2$), we can inspect the correlation between metrics—the correlation dictates the succinctness of $S : \mathcal{F} \to \mathbb{R}^k$ for satisfying improvement.

Design for Optimality goal

Theorem. For each design restriction, there exists $S : \mathcal{F} \to \mathbb{R}^k$ that satisfies the optimality objective with the following dimensionalities.

Res.	Dimensionality $k \ge$
CS	$\dim \text{aff}(\mathcal{F})$
LM	1
L	1

Future directions

- Soon: Can account for preferences for comparing metrics: $\boldsymbol{f}' \ge_{\text{Cone}(\cdot)} \boldsymbol{f}$
- Without full knowledge of $\mathcal{F}$? Say if we get samples $\boldsymbol{f}$ from set $\mathcal{F}$.
- Non-linear score design restrictions?

References

Daniel Koretz. *The Testing Charade: Pretending to Make Schools Better*. The University of Chicago Press, 2017.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.