Robust Replacement of Human Decision Makers under Distribution Shift

Dingli Wang

Department of Economics, University of California, Riverside
dwang177@ucr.edu

Job Market Talk · 2026–2027

Motivation and overview

Motivation and overview

The rule

Statistical guarantees

What the rule gives up, and the choice of threshold

Pointwise budgets

Simulations

US income data

Conclusion

When should an algorithm replace a doctor?

- Algorithms now perform as well as human experts in many prediction tasks (Esteva et al. 2017 Nature).
- Economists use them to evaluate and improve human decisions: bail hearings (Kleinberg et al. 2018 QJE), medical testing (Mullainathan & Obermeyer 2022 QJE).
- A natural next step: hand the decisions of an individual decision maker, such as a doctor, to an algorithm.

When should an algorithm replace a doctor,
before we know the cases it will see?

Measured performance can overstate performance in use

A sepsis alert

A widely used proprietary model, in an external validation: AUC **0.63**, against 0.76–0.83 reported by its developer. It missed 67% of the patients with sepsis.

Wong et al. (2021)

A change in data extraction

A hospital model validated prospectively: AUC 0.767, against 0.778 on past data. The gap came mainly from changes in how the data were accessed, extracted and transformed.

Otles et al. (2021)

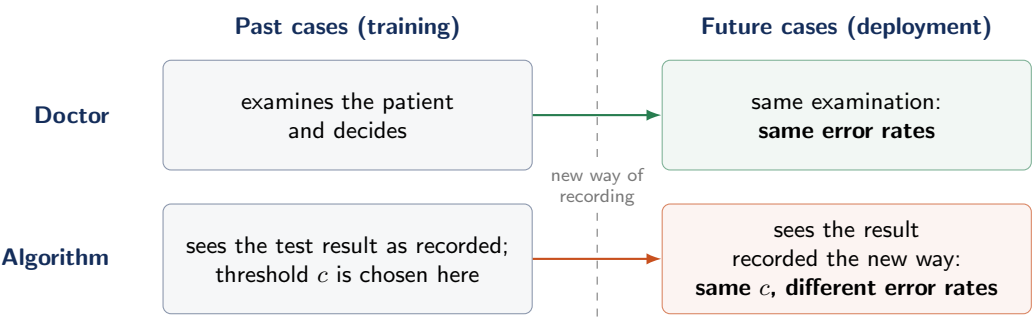
New hospitals

A model that detects pneumonia on chest radiographs generalized unevenly across three hospital systems.

Zech et al. (2018); see also Finlayson et al. (2021)

Comparisons of humans and algorithms usually test the algorithm on held-out cases from the same source as its training data.

Running example: the hospital changes how a test result is recorded



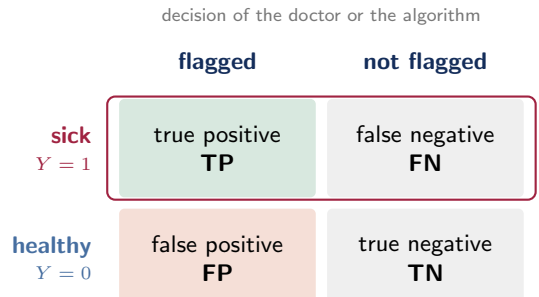
- A real case: after the ICUs in MIMIC-III changed their record-keeping software in 2008, a mortality model trained on earlier records lost 0.29 in AUROC (Nestor et al., 2019).
- The doctor's information does not change; the data seen by the algorithm do.

Two error rates

decision of the doctor or the algorithm

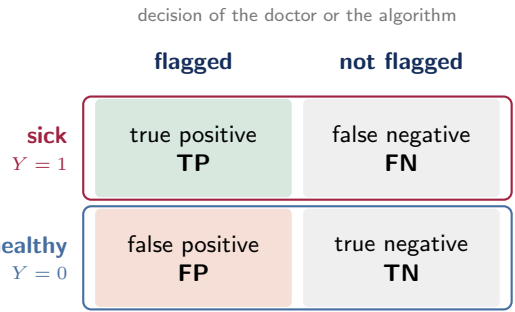
	flagged	**not flagged**
sick $Y = 1$	true positive **TP**	false negative **FN**
healthy $Y = 0$	false positive **FP**	true negative **TN**

Two error rates



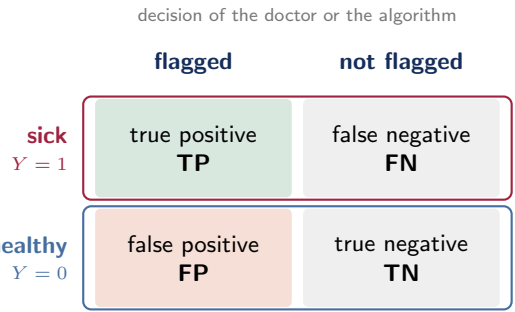
- **True positive rate**: the share of sick patients who are flagged,
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

Two error rates



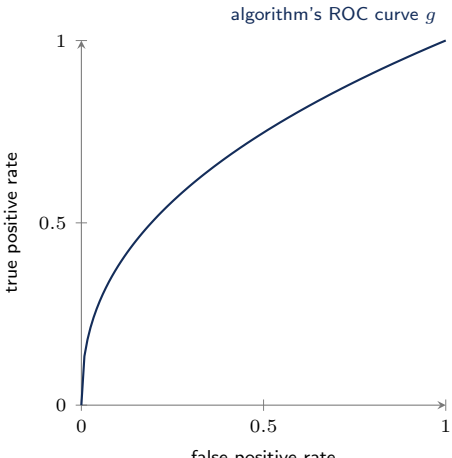
- **True positive rate**: the share of sick patients who are flagged,
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$
- **False positive rate**: the share of healthy patients who are flagged,
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}.$$

Two error rates



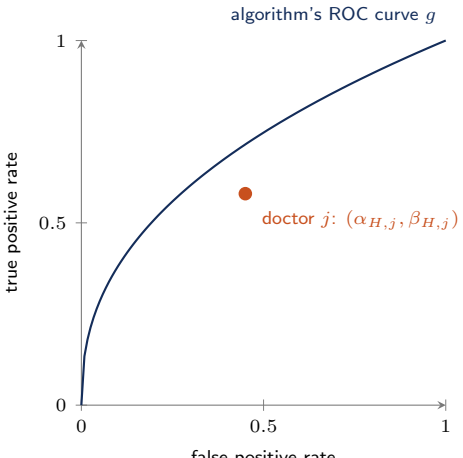
- **True positive rate**: the share of sick patients who are flagged,
  $$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$
- **False positive rate**: the share of healthy patients who are flagged,
  $$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}.$$
- Each rate uses one row only, so neither depends on how many patients are sick.
- A doctor, or the algorithm at one threshold, is one point (FPR, TPR) on the next slide.

Feng et al. (2025 MS): compare one threshold with one doctor



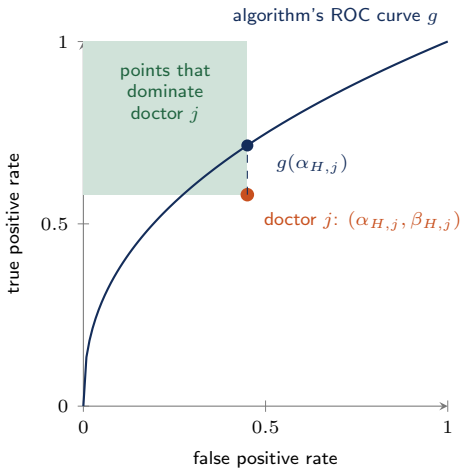
- Algorithm: score $m(X) \in [0, 1]$; flag the case if $m(X) > c$. Varying the threshold c traces its ROC curve: $g(\alpha)$ is the highest TPR at an FPR of at most α.

Feng et al. (2025 MS): compare one threshold with one doctor



- Algorithm: score $m(X) \in [0,1]$; flag the case if $m(X) > c$. Varying the threshold c traces its ROC curve: $g(\alpha)$ is the highest TPR at an FPR of at most α.
- Doctor j: a single point, the false and true positive rates $(\alpha_{H,j}, \beta_{H,j})$.

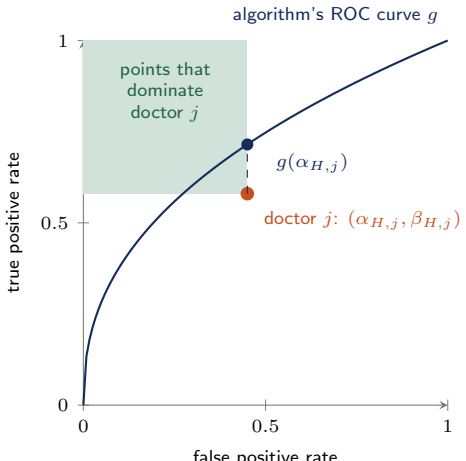
Feng et al. (2025 MS): compare one threshold with one doctor



Some point dominates the doctor iff $\beta_{H,j} \leq g(\alpha_{H,j})$.

- Algorithm: score $m(X) \in [0,1]$; flag the case if $m(X) > c$. Varying the threshold c traces its ROC curve: $g(\alpha)$ is the highest TPR at an FPR of at most α.
- Doctor j: a single point, the false and true positive rates $(\alpha_{H,j}, \beta_{H,j})$.
- A point on the curve *dominates* the doctor if its FPR is no higher and its TPR no lower: at least as good whatever the costs of the two errors.

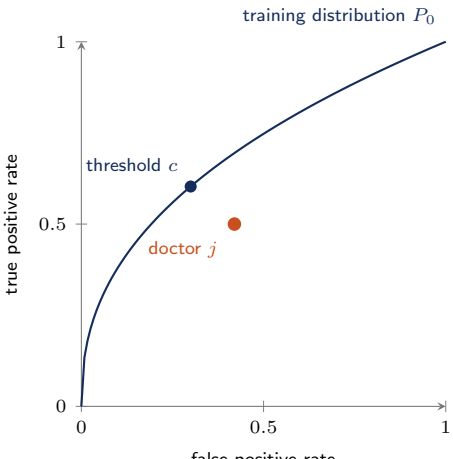
Feng et al. (2025 MS): compare one threshold with one doctor



Some point dominates the doctor iff $\beta_{H,j} \leq g(\alpha_{H,j})$.

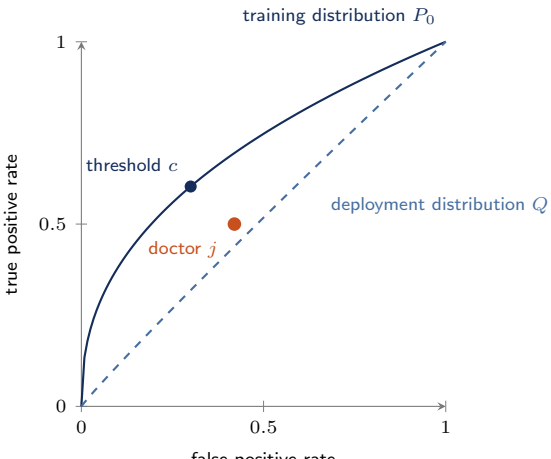
- Algorithm: score $m(X) \in [0,1]$; flag the case if $m(X) > c$. Varying the threshold c traces its ROC curve: $g(\alpha)$ is the highest TPR at an FPR of at most α.
- Doctor j: a single point, the false and true positive rates $(\alpha_{H,j}, \beta_{H,j})$.
- A point on the curve *dominates* the doctor if its FPR is no higher and its TPR no lower: at least as good whatever the costs of the two errors.
- The doctor's rates are estimated from past decisions. Replace the doctor if, for some point on the curve, the *posterior probability* that it dominates the doctor is at least τ, such as 0.95.

The curve is measured on past cases, but used on future cases



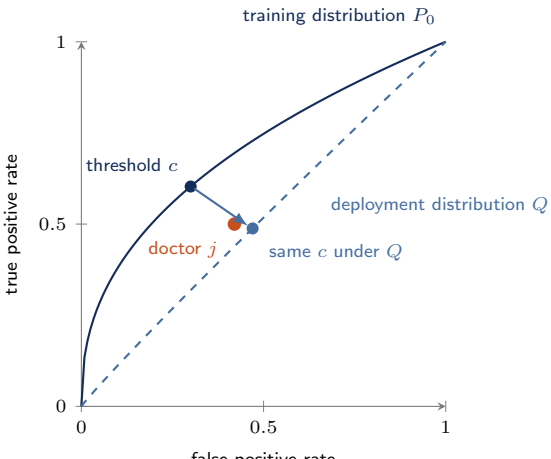
- The ROC curve is estimated on validation cases drawn from the training distribution P_0.

The curve is measured on past cases, but used on future cases



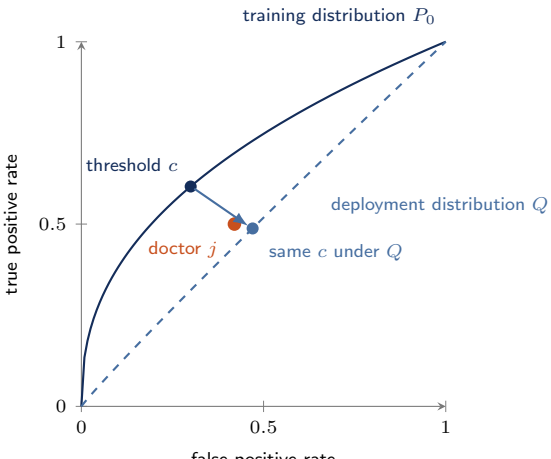
- The ROC curve is estimated on validation cases drawn from the training distribution P_0.
- At deployment, the data seen by the algorithm may change, and the curve moves.

The curve is measured on past cases, but used on future cases



- The ROC curve is estimated on validation cases drawn from the training distribution P_0.
- At deployment, the data seen by the algorithm may change, and the curve moves.
- The threshold is chosen in advance: a threshold that dominated the doctor may no longer do so.

The curve is measured on past cases, but used on future cases

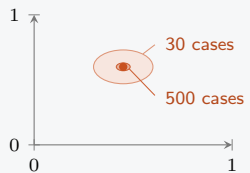


- The ROC curve is estimated on validation cases drawn from the training distribution P_0.
- At deployment, the data seen by the algorithm may change, and the curve moves.
- The threshold is chosen in advance: a threshold that dominated the doctor may no longer do so.
- The doctor's point does not move: same patients, same examination.

Two kinds of uncertainty, treated differently

The doctor's error rates

Unknown, but each past decision is informative.
⇒ a **posterior**, which tightens as the history grows.



The deployment distribution

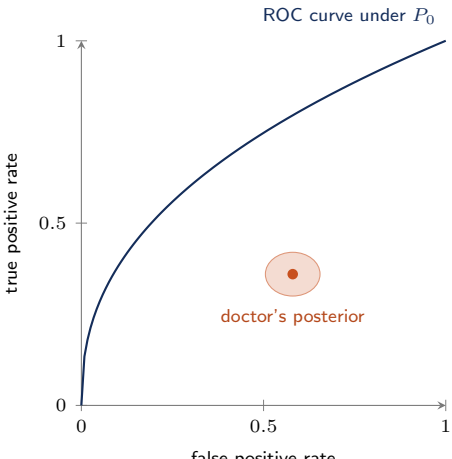
Unknown, and the past cases cannot reveal it: *all* of them come from the training distribution.

No deployment cases are observed

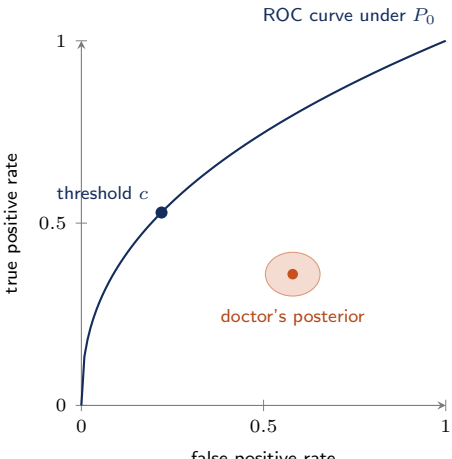
⇒ a **set** of plausible deployment distributions

The rule: at *one* threshold chosen in advance, the algorithm must dominate the doctor under *every* distribution in the set, with posterior probability at least τ.

The rule in one picture

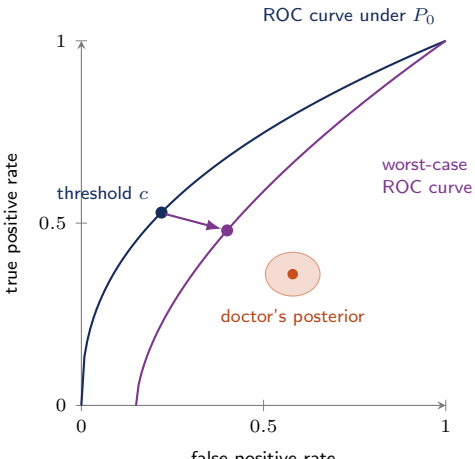


The rule in one picture



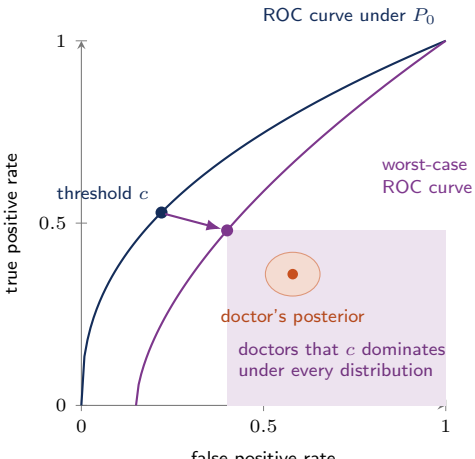
- A threshold c gives one point on the ROC curve under P_0.

The rule in one picture



- A threshold c gives one point on the ROC curve under P_0.
- Over the set, its FPR can rise and its TPR can fall. The worst case of each gives a corner; across thresholds, the corners trace the *worst-case ROC curve*.

The rule in one picture



- A threshold c gives one point on the ROC curve under P_0.
- Over the set, its FPR can rise and its TPR can fall. The worst case of each gives a corner; across thresholds, the corners trace the *worst-case ROC curve*.
- Replace the doctor if, for some c, the posterior probability that the doctor lies in the rectangle of c is at least τ.
- With a budget of zero, this is the rule of Feng et al.: the *nominal rule*.

This paper

Extends the nominal rule to an unknown deployment distribution: a posterior for each doctor's error rates, a set of plausible distributions for the algorithm's data.

1. **Guarantees.** The estimated worst-case ROC curve converges at the usual root-n rate. The rule replaces the right doctors with probability approaching one. Under every distribution in the set, it can lose little relative to retaining all doctors.

2. **The threshold matters.** The threshold with the highest posterior probability can leave a welfare loss that more data do not remove. Choosing by welfare among the thresholds that pass the test does no worse in the bound, and better in the simulations. ▸ many thresholds

3. **How the set is specified matters.** An average budget can be spent on the few cases nearest a threshold, so even a small one moves the worst-case rates a lot; a pointwise budget cannot.

Related literature

- **Humans and algorithms.** Kleinberg et al. (2018); Mullainathan & Obermeyer (2022); Feng, Hong, Tang & Wang (2025).
  → we extend their rule to an unknown deployment distribution.
- **Distributionally robust optimization.** Mohajerin Esfahani & Kuhn (2018); Blanchet & Murthy (2019); Kuhn et al. (2019); Duchi & Namkoong (2021); Gao & Kleywegt (2023); Shafieezadeh-Abadeh et al. (2015, 2019).
  → the same distances to a hyperplane, computed within each class and combined with a posterior.
- **Distribution shift.** Shimodaira (2000); Sugiyama et al. (2007); Sugiyama & Kawanabe (2012).
  → importance weighting needs the deployment distribution; here the threshold is chosen before it is seen.
- **Margin conditions, binary decisions.** Mammen & Tsybakov (1999); Audibert & Tsybakov (2007); Elliott & Lieli (2013); Feng, Hong & Nekipelov (2026).

The rule

Motivation and overview

The rule

Statistical guarantees

What the rule gives up, and the choice of threshold

Pointwise budgets

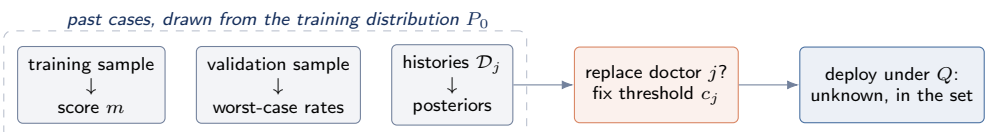
Simulations

US income data

Conclusion

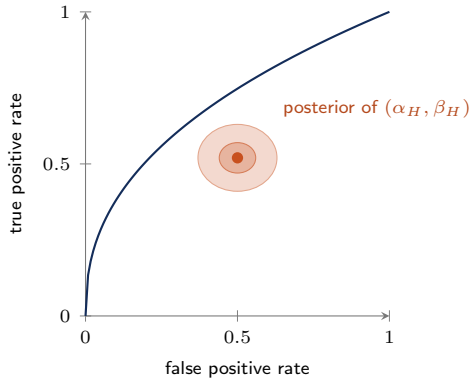
Data and timing

- J doctors. Doctor j has a history $\mathcal{D}_j$ of past cases: whether the patient was sick, and what the doctor decided.
- The algorithm's score m is fitted on a training sample and *held fixed*. A separate validation sample of n past cases is used to estimate its error rates.
- For each doctor, a planner either retains the doctor or replaces the doctor with the algorithm at *one* threshold c, chosen now and not adjusted later.



Each doctor's error rates have a posterior

- Doctor j's history: FP_j healthy patients flagged and TN_j not; TP_j sick patients flagged and FN_j not.
- The posterior for the FPR α_H is centred near $\mathrm{FP}_j/(\mathrm{FP}_j + \mathrm{TN}_j)$ and narrows as the history grows; the same holds for the TPR β_H. The two are independent.

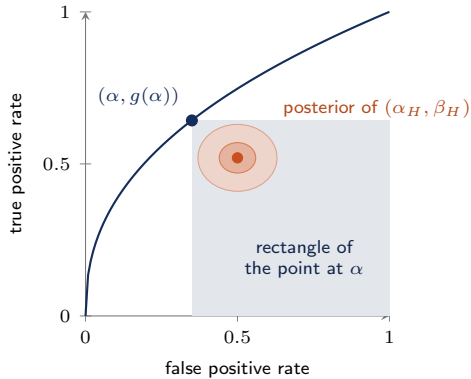


A Dirichlet prior on the four counts gives independent Beta posteriors, e.g. $\alpha_H \mid \mathcal{D}_j \sim \mathrm{Beta}(a_0 + \mathrm{FP}_j,\, a_0 + \mathrm{TN}_j)$.

Each doctor's error rates have a posterior

- Doctor j's history: FP_j healthy patients flagged and TN_j not; TP_j sick patients flagged and FN_j not.
- The posterior for the FPR α_H is centred near $\mathrm{FP}_j/(\mathrm{FP}_j + \mathrm{TN}_j)$ and narrows as the history grows; the same holds for the TPR β_H. The two are independent.
- The point $(\alpha, g(\alpha))$ on the ROC curve dominates every doctor in its rectangle: $\alpha_H \geq \alpha$ and $\beta_H \leq g(\alpha)$.

A Dirichlet prior on the four counts gives independent Beta posteriors, e.g. $\alpha_H \mid \mathcal{D}_j \sim \mathrm{Beta}(a_0 + \mathrm{FP}_j,\, a_0 + \mathrm{TN}_j)$.



What may change at deployment, and what may not

- We assume that, in every environment we consider, each doctor's error rates are those of past cases:
$$\mathbb{P}_Q(\widehat{Y}_j = 1 \mid Y = 0) = \alpha_{H,j}, \qquad \mathbb{P}_Q(\widehat{Y}_j = 1 \mid Y = 1) = \beta_{H,j}. \tag{2.4}$$
- This holds if the doctor applies the same rule h_j to information Ω_j whose distribution given Y does not change, such as what an examination shows; the distribution of the algorithm's data X given Y may change.
- Then the posterior from the history is also a posterior for the doctor's error rates on future cases.

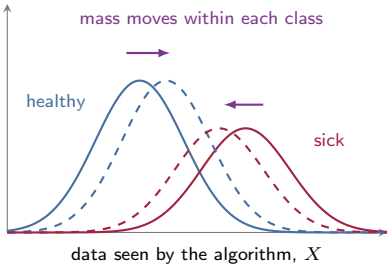
It fails when doctors rely on the same shifted records without adjusting for the change.

Measuring a shift by how far the data must move

- The size of a shift: the smallest *average* distance the algorithm's data must move to turn one distribution into the other (Wasserstein distance).
- Moving the data never changes whether a patient is sick: the prevalence is fixed, and

$$W_1(P,Q) = p\,W_1(P^1,Q^1) + (1-p)\,W_1(P^0,Q^0).$$

- The set: $\mathcal{B}_\varepsilon(P_0) = \{Q : \ W_1(Q,P_0) \le \varepsilon\}$, with *budget* ε in the units of X.



Solid: training; dashed: a deployment distribution in the set.

Worst-case rates and the worst-case ROC curve

Definition 1

For each threshold c, the worst-case TPR and FPR over the set are

$$\underline{\beta}_{\varepsilon}(c;P) = \inf_{Q\in\mathcal{B}_{\varepsilon}(P)} Q\big(m(X) > c \mid Y = 1\big), \qquad \overline{\alpha}_{\varepsilon}(c;P) = \sup_{Q\in\mathcal{B}_{\varepsilon}(P)} Q\big(m(X) > c \mid Y = 0\big),$$

and the worst-case ROC curve is $g_{\varepsilon}(\alpha;P) = \sup\big\{\underline{\beta}_{\varepsilon}(c;P) \ : \ \overline{\alpha}_{\varepsilon}(c;P) \le \alpha\big\}$.

- $g_{\varepsilon}(\alpha)$: the highest TPR that a threshold chosen in advance can guarantee, among thresholds whose FPR is at most α under every distribution in the set.
- The two worst cases are computed separately, which is what the comparison with a doctor needs: c dominates the doctor under every distribution in the set iff $\overline{\alpha}_{\varepsilon}(c) \le \alpha_H$ and $\underline{\beta}_{\varepsilon}(c) \ge \beta_H$.

Proposition 1: each worst case is a search over one number

$$\overline{\alpha}_{\varepsilon}(c) \;=\; \min_{\lambda \geq 0} \Big\{ \underbrace{\lambda \, \frac{\varepsilon}{1-p}}_{\substack{\text{price} \times \text{budget} \\ \text{per healthy case}}} + \underbrace{\mathbb{E}_P\big[\big(1 - \lambda \, d_c^{+}(X)\big)_{+} \;\big|\; Y = 0\big]}_{\text{average net gain over the healthy cases}} \Big\}$$

Blanchet & Murthy (2019) duality, applied within each class; p is the share of sick cases.

Proposition 1: each worst case is a search over one number

$$\overline{\alpha}_{\varepsilon}(c) \;=\; \min_{\lambda \geq 0} \Big\{ \underbrace{\lambda \, \frac{\varepsilon}{1-p}}_{\substack{\text{price} \times \text{budget} \\ \text{per healthy case}}} + \underbrace{\mathbb{E}_P\big[\big(1 - \lambda \, d_c^{+}(X)\big)_{+} \;\big|\; Y = 0\big]}_{\text{average net gain over the healthy cases}} \Big\}$$

- $d_c^{+}(x)$: how far a healthy case at x must move before the algorithm flags it (0 if it is flagged already). It is computed from the data, case by case; the budget caps the average of these moves.

Blanchet & Murthy (2019) duality, applied within each class; p is the share of sick cases.

Proposition 1: each worst case is a search over one number

$$\overline{\alpha}_{\varepsilon}(c) \;=\; \min_{\lambda \geq 0} \Big\{ \underbrace{\lambda \, \frac{\varepsilon}{1-p}}_{\substack{\text{price} \times \text{budget} \\ \text{per healthy case}}} + \underbrace{\mathbb{E}_P\big[\big(1 - \lambda \, d_c^{+}(X)\big)_{+} \;\big|\; Y = 0\big]}_{\text{average net gain over the healthy cases}} \Big\}$$

- $d_c^{+}(x)$: how far a healthy case at x must move before the algorithm flags it (0 if it is flagged already). It is computed from the data, case by case; the budget caps the average of these moves.
- λ: a price per unit of distance. Flagging a case gains 1 and costs λd, so it pays only when $d < 1/\lambda$: the closest cases move first.

Blanchet & Murthy (2019) duality, applied within each class; p is the share of sick cases.

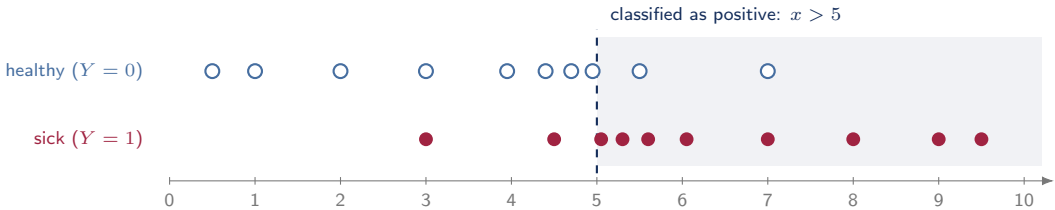
Proposition 1: each worst case is a search over one number

$$\overline{\alpha}_{\varepsilon}(c) \;=\; \min_{\lambda \geq 0} \Big\{ \quad \underbrace{\lambda \, \frac{\varepsilon}{1-p}}_{\substack{\text{price} \times \text{budget} \\ \text{per healthy case}}} \quad + \underbrace{\mathbb{E}_P\big[\big(1 - \lambda \, d_c^{+}(X)\big)_{+} \;\big|\; Y = 0\big]}_{\text{average net gain over the healthy cases}} \Big\}$$

- $d_c^{+}(x)$: how far a healthy case at x must move before the algorithm flags it (0 if it is flagged already). It is computed from the data, case by case; the budget caps the average of these moves.
- λ: a price per unit of distance. Flagging a case gains 1 and costs λd, so it pays only when $d < 1/\lambda$: the closest cases move first.
- Given p, ε and the distances, the only unknown is the number λ. The worst-case TPR has a formula of the same kind, with the distance to the region the algorithm does not flag.

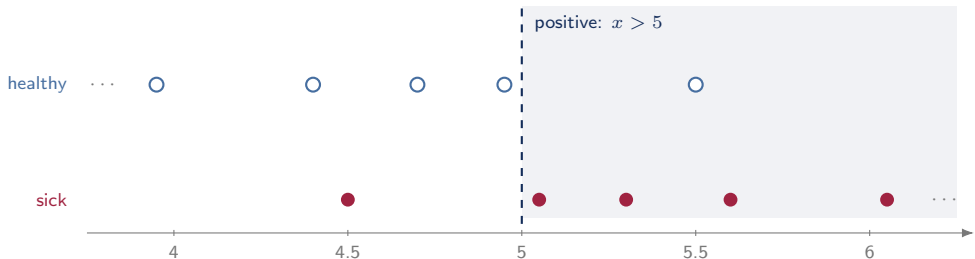
Blanchet & Murthy (2019) duality, applied within each class; p is the share of sick cases.

Example 1: twenty validation cases

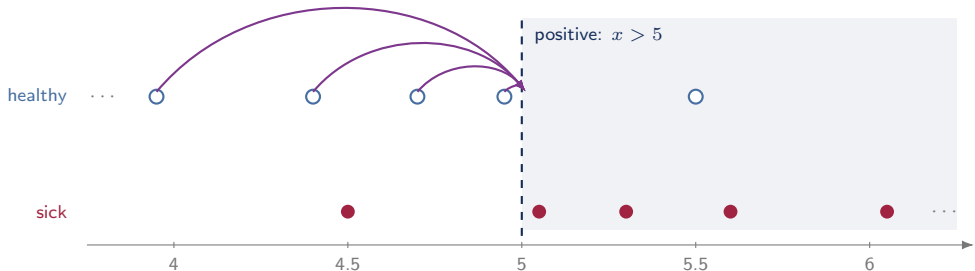


- x: one recorded value per patient, such as a test result. Score $m(x) = \{1 + \exp(5 - x)\}^{-1}$ and threshold $c = 1/2$: a case is flagged when $x > 5$.
- Two healthy and eight sick cases lie above 5: the FPR is 0.2 and the TPR is 0.8.
- The budget caps the *average* distance moved over the 20 cases. Moving one case a distance d raises the average by $d/20$, so a budget of 0.1 allows moves that add up to 2.

Example 1: an average budget of 0.1 (total distance 2)

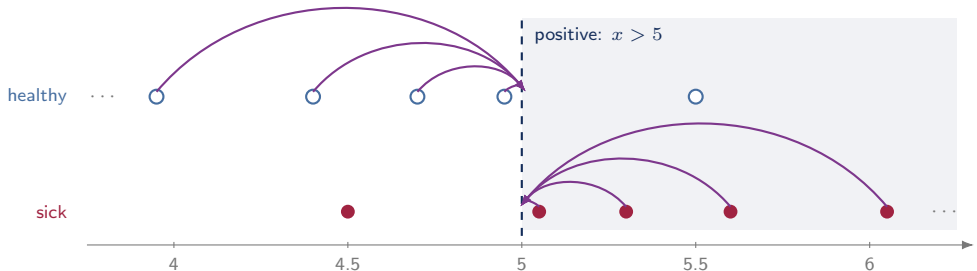


Example 1: an average budget of 0.1 (total distance 2)



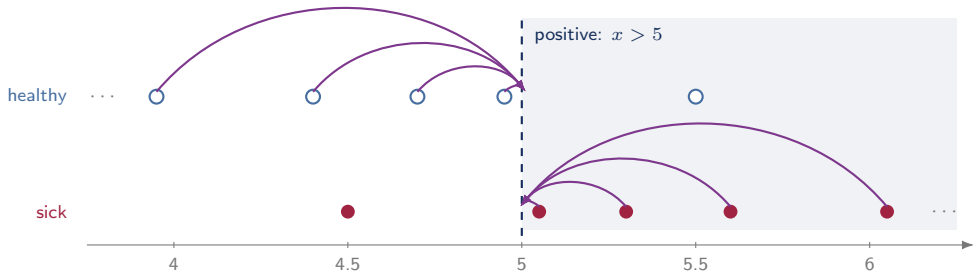
- Worst-case FPR: move the four healthy cases closest to 5 up to it ($0.05 + 0.3 + 0.6 + 1.05 = 2$). A case at exactly 5 is still negative, and the budget ends there, so the FPR approaches **0.6** but never reaches it.

Example 1: an average budget of 0.1 (total distance 2)



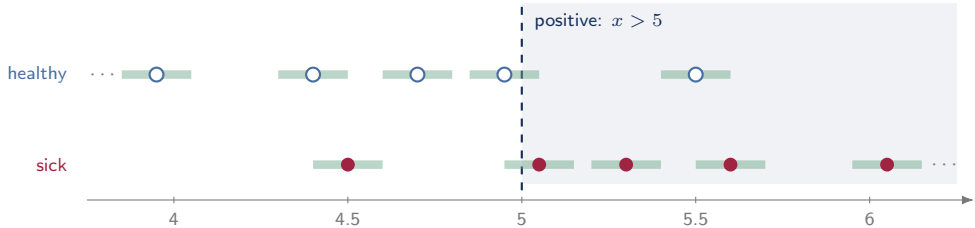
- Worst-case FPR: move the four healthy cases closest to 5 up to it ($0.05 + 0.3 + 0.6 + 1.05 = 2$). A case at exactly 5 is still negative, and the budget ends there, so the FPR approaches **0.6** but never reaches it.
- Worst-case TPR: start again and move the four sick cases just above 5 onto it. The TPR falls from 0.8 to **0.4**.

Example 1: an average budget of 0.1 (total distance 2)



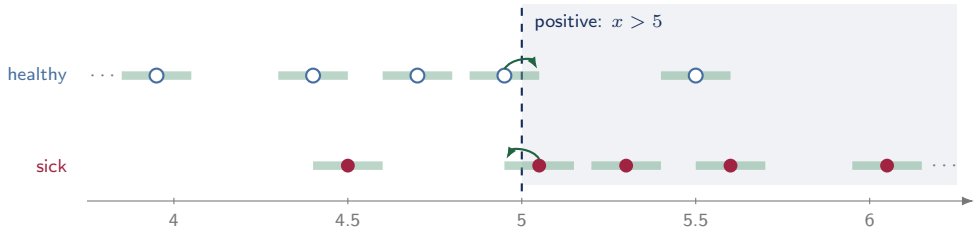
- Worst-case FPR: move the four healthy cases closest to 5 up to it ($0.05 + 0.3 + 0.6 + 1.05 = 2$). A case at exactly 5 is still negative, and the budget ends there, so the FPR approaches **0.6** but never reaches it.
- Worst-case TPR: start again and move the four sick cases just above 5 onto it. The TPR falls from 0.8 to **0.4**.
- Each worst case spends the whole budget: no single distribution in the set comes close to both.

Example 1: a pointwise budget of 0.1



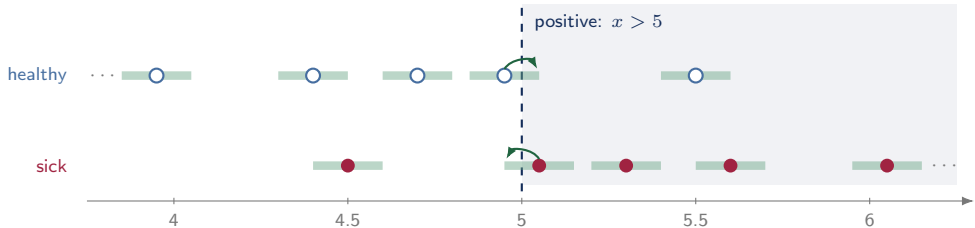
- Each case may now move by at most 0.1 (green bars).

Example 1: a pointwise budget of 0.1



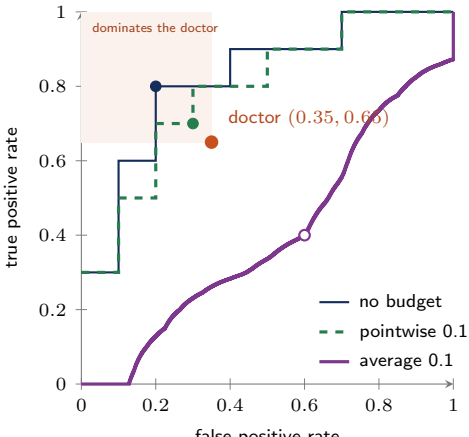
- Each case may now move by at most 0.1 (green bars).
- Only the healthy case at 4.95 and the sick case at 5.05 can cross: worst-case FPR **0.3**, worst-case TPR **0.7**.

Example 1: a pointwise budget of 0.1



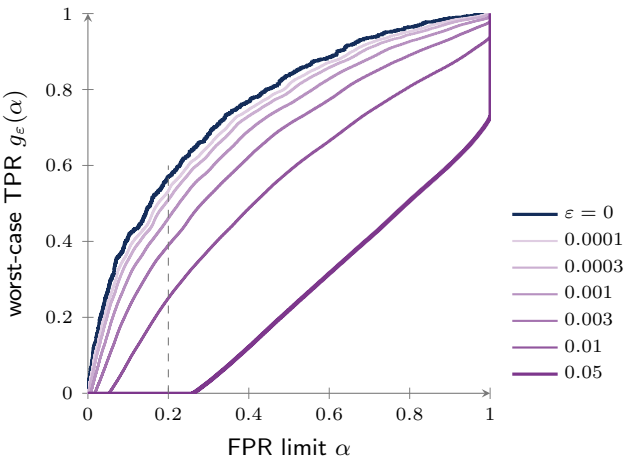
- Each case may now move by at most 0.1 (green bars).
- Only the healthy case at 4.95 and the sick case at 5.05 can cross: worst-case FPR **0.3**, worst-case TPR **0.7**.
- One distribution attains both: every healthy case up by 0.1, every sick case down by 0.1. A doctor with error rates $(0.35, 0.65)$ is dominated.

Example 1 in ROC space: same size, very different worst cases



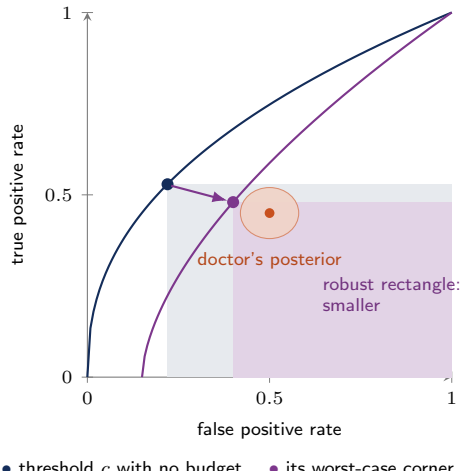
- Dots: the threshold $x > 5$ with no budget, the pointwise budget and the average budget.
- Pointwise budget 0.1: this threshold dominates the doctor.
- Average budget 0.1: no threshold does. At an FPR of 0.35, the worst-case curve guarantees a TPR of only 0.26.
- The average budget is spent on the few cases closest to each threshold.

In the simulations: small budgets already move the curve a lot



- First simulated sample (Section 5): fitted logistic score, 1,001 candidate thresholds, average budgets.
- At an FPR limit of 0.2, the TPR that can be guaranteed is 0.57 with no budget, 0.39 with $\varepsilon = 0.003$, 0.25 with $\varepsilon = 0.01$ and 0 with $\varepsilon = 0.05$.
- Why so sensitive? As in Example 1, the budget goes to the cases nearest each threshold.

The robust replacement rule



- Posterior probability that c dominates the doctor under every distribution in the set:

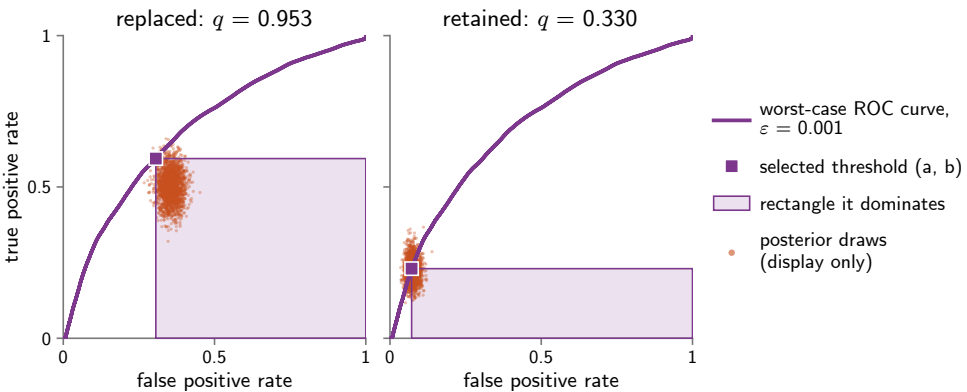
$$q_j^{\varepsilon}(c) = \Pi_j\big(\alpha_H \geq \overline{\alpha}_{\varepsilon}(c),\ \beta_H \leq \underline{\beta}_{\varepsilon}(c) \mid \mathcal{D}_j\big),$$

  a product of two Beta probabilities.

- **Definition 3**: replace doctor j if $\max_c q_j^{\varepsilon}(c) \geq \tau$, and use the maximizing threshold $\widehat{c}_j$.

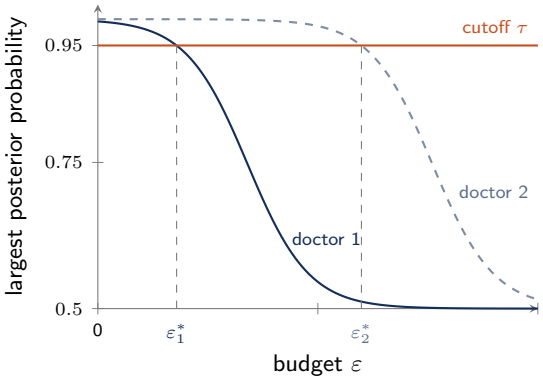
- **Proposition 2**: the robust rule replaces a subset of the doctors the nominal rule replaces.

The robust replacement rule



First simulated sample, average budget $\varepsilon = 0.001$. Dots: draws from each doctor's posterior, for display only; q is computed exactly from the two Beta posteriors.

Each replaced doctor has a critical budget



- A larger budget lowers every worst-case curve, so a doctor replaced at some budget is replaced at every smaller one.
- Each doctor the nominal rule replaces therefore has a *critical budget*: replaced below it, retained above it.
- The planner can compare it with the shifts considered plausible.

Statistical guarantees

Motivation and overview

The rule

Statistical guarantees

What the rule gives up, and the choice of threshold

Pointwise budgets

Simulations

US income data

Conclusion

Theorem 1: the worst-case ROC curve is estimated at the usual rate

Two conditions (Assumption 1)

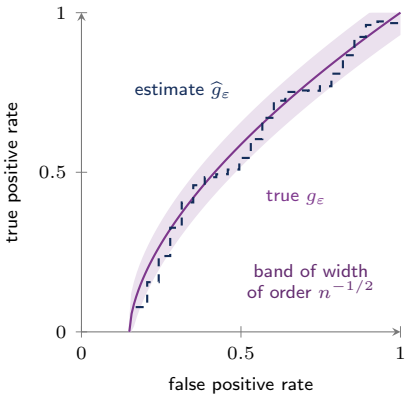
(i) The functions in the dual problems are not too rich (VC type)

(ii) On an interval $\mathcal{C}$ of thresholds, the worst-case rates are Lipschitz in c, and the slope of the worst-case FPR is bounded away from zero.

Theorem 1

$$\sup_{c\in\mathcal{C}}\left|\widehat{\underline{\beta}}_{\varepsilon}(c)-\underline{\beta}_{\varepsilon}(c)\right|+\sup_{c\in\mathcal{C}}\left|\widehat{\overline{\alpha}}_{\varepsilon}(c)-\overline{\alpha}_{\varepsilon}(c)\right|=O_p(n^{-1/2}),$$

$$\sup_{\alpha\in[\alpha_l,\alpha_u]}\left|\widehat{g}_{\varepsilon}(\alpha)-g_{\varepsilon}(\alpha)\right|=O_p(n^{-1/2}).$$



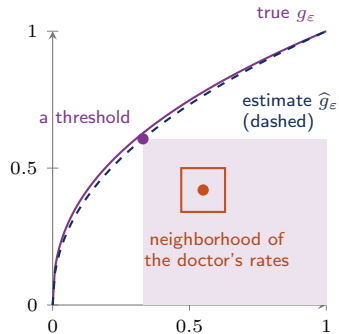
Theorem 2: the rule replaces the right doctors

- The right decision: replace the doctors below the true worst-case ROC curve, retain those above it.
- Conditions: each doctor is at least some gap away from the curve (Assumption 2); each posterior concentrates as the history grows (Assumption 3).
- Then $\mathbb{P}(\text{at least one wrong decision})$

$$\leq \underbrace{\mathbb{P}(\text{estimated curve off by} > \text{gap}/4)}_{\to 0 \text{ as the validation sample grows}}$$

$$+ \underbrace{\text{const} \times e^{-\text{const} \times \text{shortest history}}}_{\to 0 \text{ as the histories grow}}.$$

- Doctors close to the curve need more data.



Shaded: the rates that one threshold beats in the worst case. A dominated doctor's whole neighborhood fits inside, so the posterior probability tends to one.

▸ exact bound

Theorem 3: a replacement is at most slightly worse than the doctor

With sampling probability at least $1-\gamma$, for all replaced doctors j and all $Q \in \mathcal{B}_\varepsilon(P_0)$ at once,

$$\mathrm{FPR}_Q(\widehat{c}_j) \le \alpha_{H,j} + e_{\alpha,n} + t_n(\gamma), \qquad \mathrm{TPR}_Q(\widehat{c}_j) \ge \beta_{H,j} - e_{\beta,n} - t_n(\gamma).$$

- $e_{\alpha,n}, e_{\beta,n}$: the largest errors in the estimated worst-case rates over all candidate thresholds, because the threshold is chosen from the same data. $O_p(n^{-1/2})$ by Theorem 1.
- $t_n(\gamma)$: the doctor's rates are known only through the posterior. Of order $\sqrt{\log(J/\gamma)/n_{\min}}$, where $n_{\min}$ is the shortest history.
- bounds how much worse the selected threshold can be than the doctor it replaces.

Corollary 1: little to lose relative to retaining all doctors

Welfare: $\phi\,\mathrm{TPR} - \eta\,\mathrm{FPR}$, averaged over the J doctors; $\mathcal{R}_\varepsilon^{(n)}$ is the set of replaced doctors.

With sampling probability at least $1-\gamma$, for every $Q \in \mathcal{B}_\varepsilon(P_0)$ at once

$$W(\mathcal{R}_\varepsilon^{(n)}, \widehat{c}; Q) - W(\emptyset, \varnothing; Q) \;\geq\; -\frac{|\mathcal{R}_\varepsilon^{(n)}|}{J}\Big[\phi\, e_{\beta,n} + \eta\, e_{\alpha,n} + (\phi+\eta)\, t_n(\gamma)\Big].$$

- The possible loss is the share replaced times the per-doctor bound of Theorem 3. With known rates, it would be zero: every replaced doctor would be dominated under every distribution in the set.
- For the nominal rule, the set contains only P_0: its guarantee covers only the training distribution.

What the rule gives up, and the choice of threshold

Motivation and overview

The rule

Statistical guarantees

What the rule gives up, and the choice of threshold

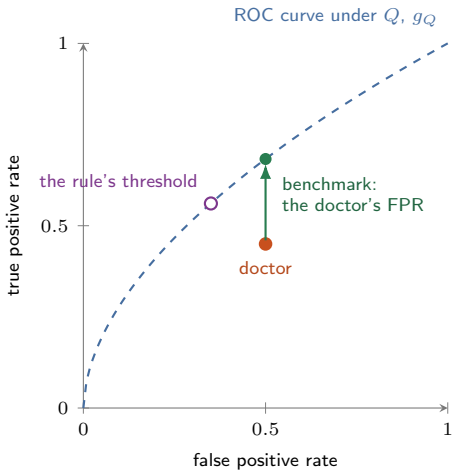
Pointwise budgets

Simulations

US income data

Conclusion

A benchmark that knows the deployment distribution



- The *matched-FPR benchmark* knows Q and every doctor's error rates, so it cannot be implemented.
- It replaces the doctors below g_Q, at the threshold with the doctor's FPR: no more false positives, as many true positives as possible.
- Regret = benchmark's welfare − the rule's welfare. It splits into common, missed and extra replacements.

Theorem 4: what the rule gives up has three kinds of terms

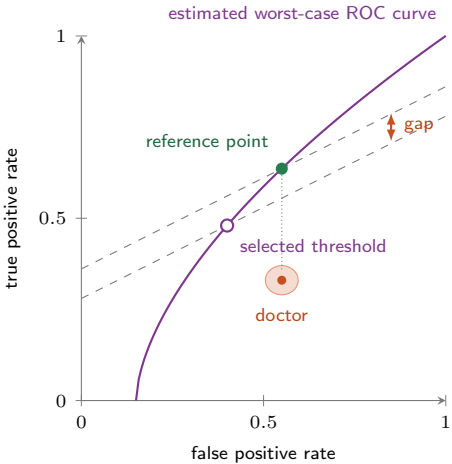
With sampling probability at least $1-\gamma$, for every Q in the set at which the benchmark is defined:

$$\mathrm{Regret}(\varepsilon, Q) \;\le\; \underbrace{\widehat{K}_n}_{\text{threshold term}} \;+\; \underbrace{\widehat{\pi}_n\{\phi\Delta_\varepsilon(Q)+v_n\}}_{\text{common replacements}} \;+\; \underbrace{\phi\, C_M\{\Delta_\varepsilon(Q)+h_n\}^{1+\gamma_M}}_{\text{missed}} \;+\; \underbrace{C_M\, u_n\, h_n^{\gamma_M}}_{\text{extra}}$$

- $\Delta_\varepsilon(Q)$: how far the worst-case ROC curve lies below g_Q, the *price of guarding against shifts*.
- h_n, u_n, v_n: *estimation* terms; they vanish as the validation sample and the histories grow.
- $\widehat{K}_n$: the *threshold term*, computed from the data. It need not vanish.

Missed and extra replacements are bounded with a margin condition: few doctors lie close to g_Q (Assumption 4).

What the threshold term measures



- Reference point: the estimated worst-case curve at the doctor's posterior-median FPR.
- Dashed lines: equal welfare $\phi\,\mathrm{TPR} - \eta\,\mathrm{FPR}$ (here $\phi = 2$, $\eta = 1$).
- The doctor's threshold term is the welfare gap between the two points, if positive; $\widehat{K}_n$ adds it up over the replaced doctors and divides by the number of doctors J.
- The rule maximizes the posterior probability, not welfare, so this term need not vanish, however much data we have.

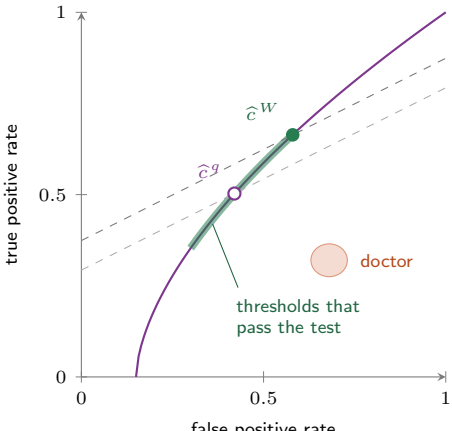
Any threshold that passes the test solves the replacement problem

The replacement decision as a decision problem

$$\max_{r_j \in \{0,1\},\ c} \ r_j \qquad \text{subject to} \qquad r_j \left\{1 - q_j^{\varepsilon}(c)\right\} \leq 1 - \tau.$$

- Replacing frees the doctor for other work, but only if some threshold passes the posterior test $q_j^{\varepsilon}(c) \geq \tau$.
- Retaining always satisfies the constraint; replacing satisfies it with *any* threshold that passes the test.
- The problem does not say which of these thresholds to use. Definition 3 takes the one with the highest posterior probability.

The welfare choice



- Among the thresholds that pass the test, take the one with the highest welfare lower bound $\widehat{V}_\varepsilon(c) = \phi\,\widehat{\beta}_\varepsilon(c) - \eta\,\widehat{\alpha}_\varepsilon(c)$: the *welfare choice* $\widehat{c}^W$.
- Same doctors replaced; only the threshold changes, and $\widehat{K}_n(\widehat{c}^W) \leq \widehat{K}_n(\widehat{c}^q)$.
- Theorems 2–3, Corollary 1 and Theorem 4 also hold for $\widehat{c}^W$. It needs the error costs (ϕ, η); $\widehat{c}^q$ does not.

Pointwise budgets

Motivation and overview

The rule

Statistical guarantees

What the rule gives up, and the choice of threshold

Pointwise budgets

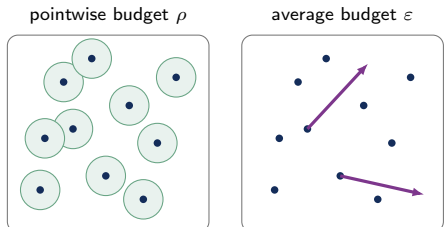
Simulations

US income data

Conclusion

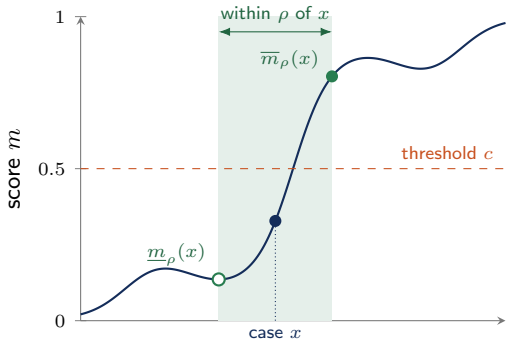
A pointwise budget: every case moves at most ρ

- An average budget limits movement only on average: a few cases may move far if the rest move little.
- A pointwise budget ρ limits how far *each* case can move. It fits shifts that move each case by at most a known distance, such as a change in measurement equipment.
- For $\rho \leq \varepsilon$, the set $\tilde{\mathcal{B}}_\rho(P)$ lies inside $\mathcal{B}_\varepsilon(P)$. With $\rho = \varepsilon$, both allow the same average movement, but only the average budget can concentrate it on a few cases.



Left: every case moves at most ρ. Right: the movement may concentrate on a few cases.

The worst cases come from the lowest and highest scores nearby



For $\rho \le \varepsilon$: $g_\varepsilon \le \tilde{g}_\rho \le g$. A pointwise budget is milder than an average budget of the same size.

- For each case, take the lowest and highest scores within distance ρ: $\underline{m}_\rho(x)$ and $\overline{m}_\rho(x)$.
- A healthy case can become a false positive iff $\overline{m}_\rho(x) > c$; a sick case can be missed iff $\underline{m}_\rho(x) \le c$.
- **Proposition 3**: so the worst-case rates are exactly

$$\tilde{\alpha}_\rho(c) = \mathbb{P}\big(\overline{m}_\rho(X) > c \mid Y = 0\big),$$
$$\tilde{\beta}_\rho(c) = \mathbb{P}\big(\underline{m}_\rho(X) > c \mid Y = 1\big).$$

Simulations

Motivation and overview

The rule

Statistical guarantees

What the rule gives up, and the choice of threshold

Pointwise budgets

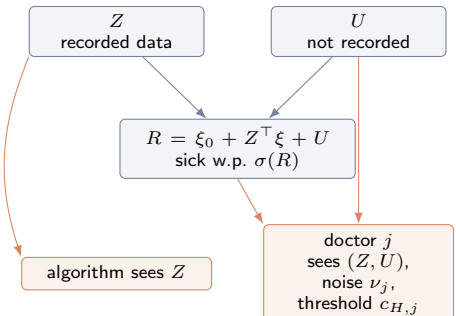
Simulations

US income data

Conclusion

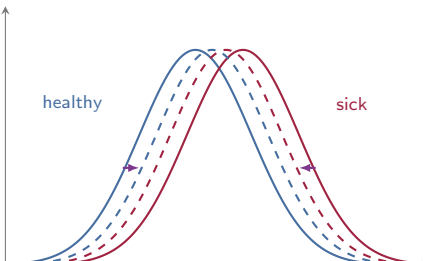
Design: the algorithm's data shift, the doctors' information does not

- $Z \sim N(0, I_5)$ and $U \sim N(0, \sigma_u^2)$, independent. With $R = \xi_0 + Z^\top \xi + U$, the patient is sick with probability $\sigma(R)$.
- The algorithm sees $X = Z$. Doctor j also sees U, information that is not recorded, and decides $\widehat{Y}_j = \mathbb{I}\{\sigma(R) + \nu_j > c_{H,j}\}$.
- The noise ν_j makes doctors less accurate than their information allows, so the algorithm can dominate some of them.



$\xi = (1, 0.5, -0.3, 0.2, -0.1)^\top$, $\xi_0 = -1.5$, $\sigma_u = 0.15$, $\sigma_\nu = 0.225$; $c_{H,j} = 0.15 + 0.7 V_j$, $V_j \sim \text{Beta}(2, 3)$.

At deployment, the two classes move toward each other



Schematic, $\delta = 0.3$. Solid: training; dashed: deployment.

- Deployment distribution Q_δ, with $\delta \geq 0$ the *distance*: the algorithm now sees Z with its first variable recorded as $Z_1 + \delta$ for healthy patients and $Z_1 - \delta$ for sick ones, as after a change in how it is recorded.
- The doctors still examine the patient and see (Z, U), so their error rates do not change.
- Only the distribution of X within each class moves; the prevalence does not.
- $\mathcal{W}_\kappa(P_0, Q_\delta) = \delta$ exactly, so Q_δ is in the set if and only if $\delta \leq \varepsilon$ (or $\delta \leq \rho$).

Implementation

Each of 100 replications

- 10,000 training cases $\rightarrow$ ℓ_2-penalized logistic score, then held fixed
- 5,000 validation cases $\rightarrow$ worst-case rates at 1,001 thresholds
- 500 doctors, each with 500 past cases $\rightarrow$ Beta posteriors ($a_0 = 0.01$)
- replace if the posterior probability is at least $\tau = 0.95$

Budgets and evaluation

- average budgets $0,\ 0.05,\ 0.1,\ 0.2,\ 0.3$, and a finer grid below 0.05
- pointwise budgets $0,\ 0.05,\ 0.1,\ 0.2,\ 0.3$
- deployment distances $\delta \in \{0,\ 0.025,\ 0.05,\ 0.1,\ 0.3,\ 0.5,\ 1\}$
- welfare $2\,\mathrm{TPR} - \mathrm{FPR}$; also $(\phi, \eta) = (1, 1)$ and $(5, 1)$

- Error rates are computed by numerical integration, not from a test sample; the benchmark uses the same rates.
- The thresholds are chosen before, and without, any deployment case.

Average budgets of 0.05–0.3 retain every doctor

δ	Nominal welfare	Robust welfare	Robust − nominal (s.e.)	Harmful among nominal repl. (%)
0	0.7323	0.6498	-0.0826 (0.0017)	0.17
0.025	0.7225	0.6498	-0.0727 (0.0015)	0.31
0.05	0.7126	0.6498	-0.0628 (0.0013)	0.70
0.1	0.6926	0.6498	-0.0428 (0.0009)	2.82
0.3	0.6105	0.6498	$+0.0392$ (0.0010)	99.94
0.5	0.5270	0.6498	$+0.1228$ (0.0027)	100
1	0.3272	0.6498	$+0.3226$ (0.0068)	100

Table 1. Means over 100 replications, $(\phi, \eta) = (2, 1)$; robust: $\varepsilon = 0.05$ (0.1, 0.2 and 0.3 give the same decisions).

- Nominal rule: replaces 41.51% of the doctors.
- Average budgets: nobody is replaced. At $\varepsilon = 0.05$ the largest posterior probability is 5.85×10^{-10}.

Average budgets of 0.05–0.3 retain every doctor

δ	Nominal welfare	Robust welfare	Robust − nominal (s.e.)	Harmful among nominal repl. (%)
0	0.7323	0.6498	-0.0826 (0.0017)	0.17
0.025	0.7225	0.6498	-0.0727 (0.0015)	0.31
0.05	0.7126	0.6498	-0.0628 (0.0013)	0.70
0.1	0.6926	0.6498	-0.0428 (0.0009)	2.82
0.3	0.6105	0.6498	$+0.0392$ (0.0010)	99.94
0.5	0.5270	0.6498	$+0.1228$ (0.0027)	100
1	0.3272	0.6498	$+0.3226$ (0.0068)	100

Table 1. Means over 100 replications, $(\phi, \eta) = (2, 1)$; robust: $\varepsilon = 0.05$ (0.1, 0.2 and 0.3 give the same decisions).

- Nominal rule: replaces 41.51% of the doctors.
- Average budgets: nobody is replaced. At $\varepsilon = 0.05$ the largest posterior probability is 5.85×10^{-10}.

- Inside the set ($\delta \leq 0.05$): the robust rule gives up the nominal rule's gains.

Average budgets of 0.05–0.3 retain every doctor

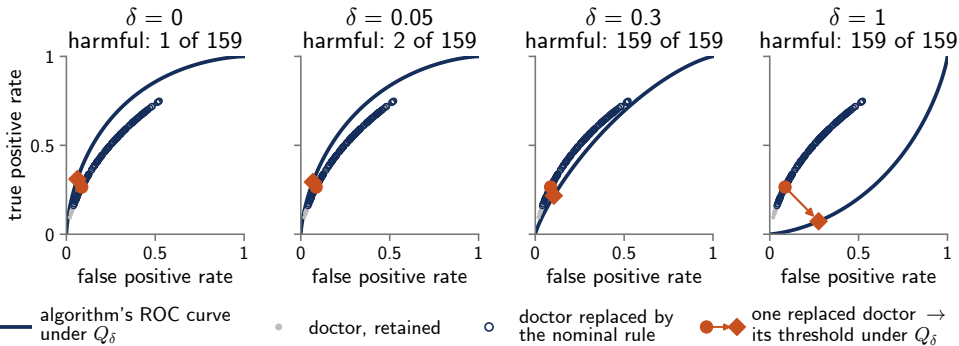
δ	Nominal welfare	Robust welfare	Robust − nominal (s.e.)	Harmful among nominal repl. (%)
0	0.7323	0.6498	−0.0826 (0.0017)	0.17
0.025	0.7225	0.6498	−0.0727 (0.0015)	0.31
0.05	0.7126	0.6498	−0.0628 (0.0013)	0.70
0.1	0.6926	0.6498	−0.0428 (0.0009)	2.82
0.3	0.6105	0.6498	+0.0392 (0.0010)	99.94
0.5	0.5270	0.6498	+0.1228 (0.0027)	100
1	0.3272	0.6498	+0.3226 (0.0068)	100

Table 1. Means over 100 replications, $(\phi, \eta) = (2, 1)$; robust: $\varepsilon = 0.05$ (0.1, 0.2 and 0.3 give the same decisions).

- Nominal rule: replaces 41.51% of the doctors.
- Average budgets: nobody is replaced. At $\varepsilon = 0.05$ the largest posterior probability is 5.85×10^{-10}.

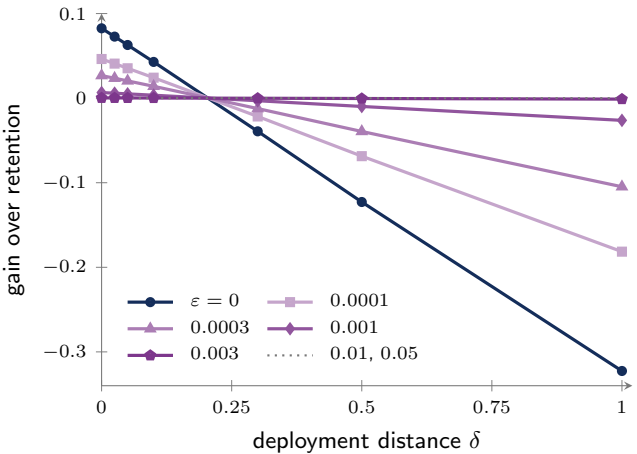
- Inside the set ($\delta \leq 0.05$): the robust rule gives up the nominal rule's gains.
- After large shifts: it avoids the nominal rule's losses, which keep growing after every replacement has become harmful.

Why the nominal rule's replacements turn harmful



- The doctors' points and the chosen thresholds stay where they are; the algorithm's ROC curve falls as δ grows.
- Harmful share among the nominal rule's replacements, over 100 replications: 0.17% at $\delta = 0$, 0.70% at 0.05, 99.94% at 0.3 and 100% at 1.

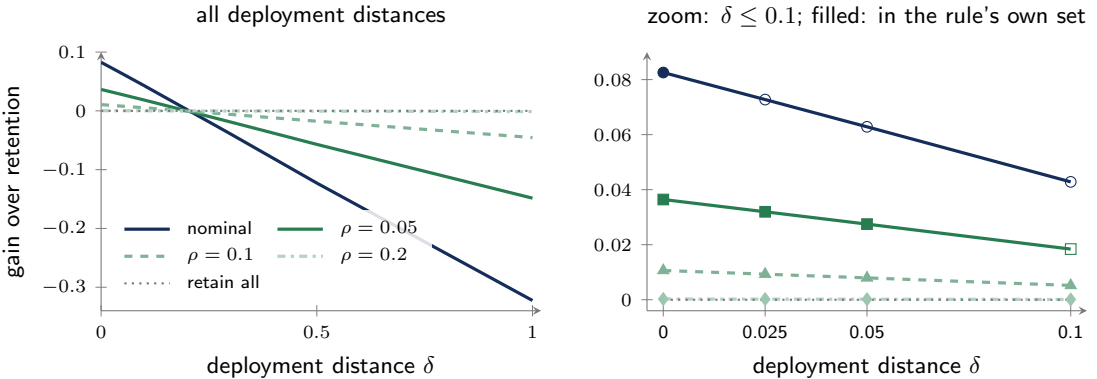
Smaller average budgets: a gradual transition



budget ε	replaced
0	41.51%
0.0001	22.79%
0.0003	12.98%
0.001	3.18%
0.003	0.14%
0.01, 0.05	0

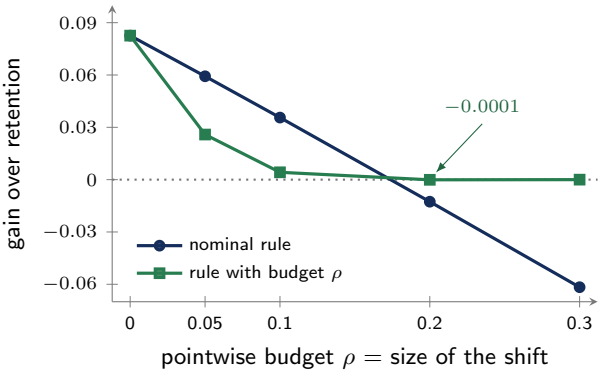
- Gains and losses both shrink toward zero.
- In the uniform example, $\varepsilon = 0.003$ already moves both rates by 0.11.
- At $\delta = \varepsilon$, the edge of each set: within 0.00004 of $\delta = 0$.

Pointwise budgets: some doctors replaced, gains inside the set



- $\rho = 0.05$ replaces 18.68% of the doctors (gain 0.0275 at $\delta = 0.05$); $\rho = 0.1$ replaces 5.61% (0.0052 at $\delta = 0.1$); $\rho = 0.3$ replaces nobody.
- After large shifts, $\rho = 0.05$ loses about half as much as the nominal rule: 0.1485 against 0.3226 at $\delta = 1$.

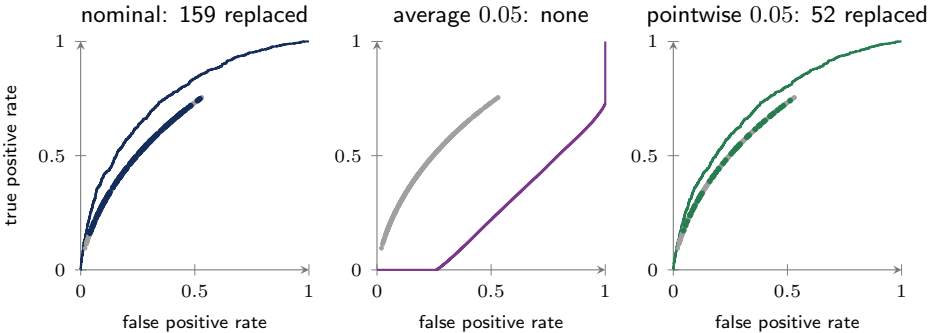
At the worst case in its own set, each pointwise rule loses at most 0.0001



Means over 100 replications; welfare $2\,\mathrm{TPR} - \mathrm{FPR}$; threshold $\widehat{c}^{\,q}$.

- Each rule meets the worst case in its own set: healthy cases move ρ toward higher scores, sick cases ρ toward lower ones.
- $\rho = 0.05$ and 0.1: gains of 0.026 and 0.004.
- $\rho = 0.2$: 0.15% of the doctors are replaced and 95% of them are harmful, but the loss is 0.0001. The nominal rule loses 0.013.
- $\rho = 0.3$: nobody is replaced; the nominal rule loses 0.062.

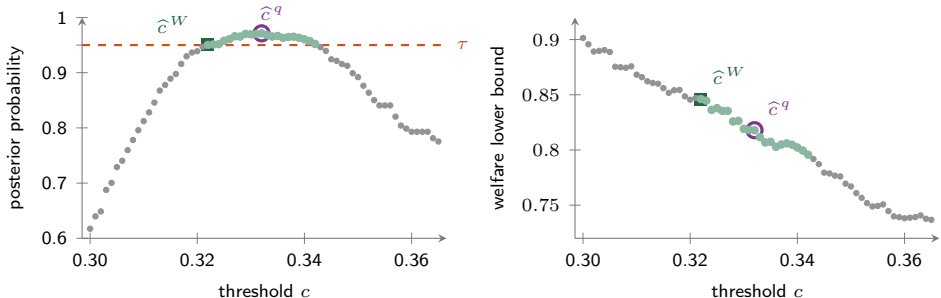
Same size, different budgets: 0.05



First replication. Curves: estimated worst-case ROC curves. Points: the doctors' true error rates; colored if the rule replaces the doctor, which it decides from the doctor's history.

- Over 100 replications: nominal 41.51%, average 0, pointwise 18.68%; in every replication the three sets of replaced doctors are nested.
- The reason is the one in Example 1: an average budget is spent on the few cases nearest the threshold.

One doctor: the welfare choice sits at the edge of the passing set



Pointwise budget 0.05, weights $(2, 1)$. Green: the 21 thresholds that pass the test; gray: those that do not.

- $\widehat{c}^q = 0.332$: posterior probability 0.971, welfare lower bound 0.818.
- $\widehat{c}^W = 0.322$: 0.950 and 0.846. The same doctor is replaced, at a lower threshold.

The welfare choice trades posterior probability for welfare

		posterior probability		welfare lower bound		$10^3\,\widehat{K}_n$	
ρ	replaced	$\widehat{c}^{\,q}$	$\widehat{c}^{\,W}$	$\widehat{c}^{\,q}$	$\widehat{c}^{\,W}$	$\widehat{c}^{\,q}$	$\widehat{c}^{\,W}$
0	41.51%	0.9824	0.9570	0.9335	0.9605	32.73	21.34
0.05	18.68%	0.9774	0.9564	0.9003	0.9207	13.06	9.15
0.1	5.61%	0.9732	0.9558	0.8574	0.8736	3.58	2.64
0.2	0.15%	0.9661	0.9548	0.7660	0.7764	0.08	0.07

Table 2. Pointwise budgets; $\tau = 0.95$; weights $(2, 1)$. Posterior probabilities and welfare lower bounds: averages over the replaced doctors, then over the replications with replacements. Shares replaced and $\widehat{K}_n$: averages over all replications and doctors.

- Both choices replace the same doctors and pass the same test.
- At $\rho = 0.05$: the posterior probability falls from 0.977 to 0.956, the welfare lower bound rises from 0.900 to 0.921, and the threshold term falls by about 30%.

US income data

Motivation and overview

The rule

Statistical guarantees

What the rule gives up, and the choice of threshold

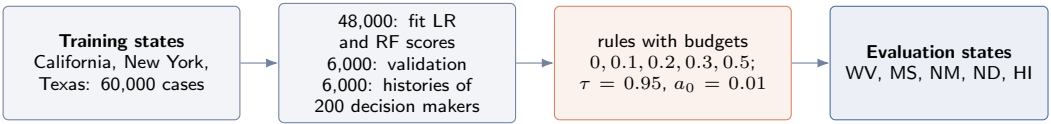
Pointwise budgets

Simulations

US income data

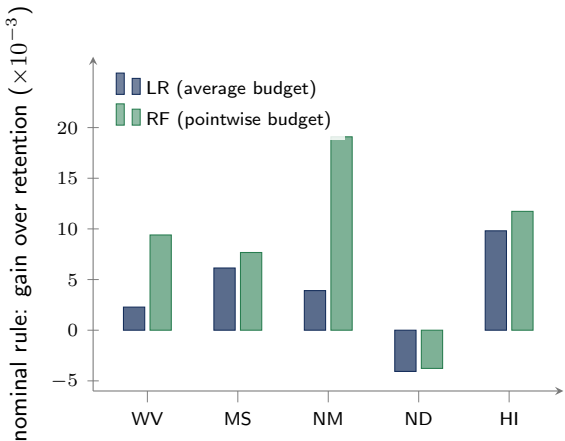
Conclusion

US income data with synthetic decision makers



- ACSIncome (Ding et al. 2021): predict whether annual income exceeds \$50,000 from ten variables.
- LR: average budget, exact worst-case rates. RF (100 trees): pointwise budget, search with $K = 200$ proposals.
- Synthetic decision makers: the LR score plus noise independent of income, against their own thresholds; about 30 past cases each, later extended to 150 and 300.

The nominal rule changes welfare slightly; budgets retain everyone



- The nominal rule replaces 5 of the 200 decision makers with the LR and 10 with the RF: it gains in four states and loses in North Dakota.
- With every budget from 0.1 to 0.5, nobody is replaced. At 0.1 the largest posterior probability is 0.0374 (LR) and 0.8954 (RF).
- So the budgeted rules give up four small gains and avoid one small loss.

Table 3. Welfare $2\,\mathrm{TPR} - \mathrm{FPR}$; one sample split, no standard errors.

Longer histories: more replacements, but the budgets still bind

	LR, average budget			LR, pointwise budget			RF, pointwise budget		
cases per decision maker	0	0.001	0.003	0.05	0.1	0.2	0	0.05	0.1
about 30 (original)	5	1	0	3	2	0	10	1	0
150	27	3	0	16	3	0	60	4	0
300	66	9	0	26	11	1	139	13	0

Decision makers replaced, of 200. Budget 0 is the nominal rule (the same for both LR columns). Larger budgets replace nobody: LR average 0.01–0.5, LR pointwise 0.3–0.5, RF 0.1–0.5. Histories nested, one realization each; RF worst cases by a search with $K = 200$.

- Longer histories: more decision makers are replaced. With 300 cases each, the nominal rule replaces 66 (LR) and 139 (RF) of the 200.
- The budgets still bind: an average budget of 0.003 replaces nobody, at any history length; a pointwise budget of 0.1 replaces at most 11.
- So the short histories explain part of the result, not all of it.

Conclusion

Motivation and overview

The rule

Statistical guarantees

What the rule gives up, and the choice of threshold

Pointwise budgets

Simulations

US income data

Conclusion

Summary

1. **A rule for an unknown deployment distribution.** Replace a doctor only if, at one threshold chosen in advance, the algorithm dominates the doctor under every distribution in a set around the training distribution, with high posterior probability. If the doctors' error rates do not change, the possible loss relative to retaining all doctors shrinks as the samples grow.

2. **The threshold still has to be chosen.** The threshold with the highest posterior probability need not have the highest welfare, even without estimation error. With error costs, the welfare choice replaces the same doctors, and its threshold term is no larger.

3. **So does the budget.** An average budget can be spent on the few cases nearest a threshold. In the simulations, average budgets of 0.05–0.3 retain every doctor; pointwise budgets of 0.05 and 0.1 replace some and raise mean welfare at every evaluated distance inside the set, including the worst case.

Guidance for practice

- Report the decisions over a grid of budgets, and each replaced doctor's critical budget.
- A budget is not a distance between two observed samples: samples with different prevalences are infinitely far apart, and a distance that pools the two classes is never larger than the label-preserving one.
- Use a pointwise budget only when each case can move by at most a known distance, as after a change in measurement equipment; otherwise use an average budget.
- Choose the threshold by welfare when the costs of the two errors can be stated; otherwise the threshold with the highest posterior probability still answers whom to replace.

Thank you

Robust Replacement of Human Decision Makers under Distribution Shift

Dingli Wang · University of California, Riverside · dwang177@ucr.edu

Backup slides

Backup: Proposition 1, an exact dual for the worst-case rates

Blanchet & Murthy (2019) duality, applied within each class

Let $d_c^-(x)$ be the distance from x to $\{m \le c\}$ and $d_c^+(x)$ the distance to $\{m > c\}$. For $\varepsilon > 0$,

$$\underline{\beta}_\varepsilon(c;P) = \sup_{\lambda \ge 0} \Big\{ \tfrac{1}{p}\, \mathbb{E}_P\big[Y \min\{1, \lambda d_c^-(X)\}\big] - \tfrac{\lambda}{p}\, \varepsilon \Big\},$$

$$\overline{\alpha}_\varepsilon(c;P) = \inf_{\lambda \ge 0} \Big\{ \tfrac{\lambda}{1-p}\, \varepsilon + \tfrac{1}{1-p}\, \mathbb{E}_P\big[(1-Y)\,(1-\lambda d_c^+(X))_+\big] \Big\},$$

and both optimizations can be restricted to $0 \le \lambda \le 1/\varepsilon$.

- λ is a price per unit of distance moved: a case is moved exactly when its distance is below $1/\lambda$, so the closest cases move first.
- Logistic score $m(x) = \sigma(w_0 + x^\top w)$, $k_c = \log\{c/(1-c)\}$: $d_c^-(x) = (z(x) - k_c)_+ / \|w\|_2$ and $d_c^+(x) = (k_c - z(x))_+ / \|w\|_2$, the distances to a half-space (Shafieezadeh-Abadeh et al. 2015, 2019), here class by class.

Backup: Proposition 1, proof sketch

- A coupling with finite label-preserving cost moves mass only within a class, so the class masses are p and $1-p$ under both marginals, and $\mathcal{W}_\kappa = pW_1(P^1, Q^1) + (1-p)W_1(P^0, Q^0)$.
- Hence $\underline{\beta}_\varepsilon(c; P) = \inf_{W_1(\nu, P^1) \le \varepsilon/p} \nu(S_c)$ and $\overline{\alpha}_\varepsilon(c; P) = \sup_{W_1(\nu, P^0) \le \varepsilon/(1-p)} \nu(S_c)$, with $S_c = \{m > c\}$.
- TPR: $\{m \le c\}$ is closed; apply Blanchet & Murthy (2019, Thm. 1) to its indicator and subtract from one.
- FPR: S_c is open. Lemma 1: for $r > 0$, $\sup_{W_1(\nu,\mu) \le r} \nu(A) = \sup_{W_1(\nu,\mu) \le r} \nu(\bar{A})$. Mix ν with μ to free a little budget, then push the mass on $\bar{A}$ a short distance into A.
- $\lambda \le 1/\varepsilon$: multiplied by p and $1-p$, the TPR objective is at most $p - \lambda\varepsilon$, and the FPR objective at least $\lambda\varepsilon$.

Backup: why the rates at budget zero are computed directly

- At $\varepsilon = 0$ the FPR dual gives $P^0(\overline{S_c})$, the probability of the *closure* of $S_c = \{m > c\}$.
- Example: $m(x) = (1 + e^{-x})^{-1}$, $c = 1/2$, all healthy cases at 0. The strict rule classifies them as negative, so the FPR is 0, but the dual at budget zero returns 1.
- At every positive budget the worst-case FPR is indeed 1: all healthy cases can be moved an arbitrarily small distance to the right of zero.
- So the nominal rates and the constant rules are computed from the definition, and the dual is used only for $\varepsilon > 0$ with two nonempty decision regions.

Backup: Lemma 3, distances, complexity and computation for logistic scores

- Distances: the projection onto the hyperplane $\{z = k_c\}$ is the closest point of the half-space; Cauchy–Schwarz rules out anything shorter.
- Complexity: within a class, the dual functions are $\ell(az + d)$ with $\ell(v) = \min\{1, \max\{0, v\}\}$ and $a = \lambda/\|w\|_2$. Their subgraphs are half-planes in (z, u), so the classes are VC-subgraph, which gives Assumption 1(i).
- Computation: with distances d_i and $V(d, \varepsilon) = \max\{\frac{1}{n}\sum u_i : \ 0 \le u_i \le 1, \ \frac{1}{n}\sum d_i u_i \le \varepsilon\}$,

$$\underline{\widehat{\beta}}_{\varepsilon}(c) = 1 - \frac{V((d_c^-(X_i))_{Y_i=1}, \varepsilon)}{n_1/n}, \qquad \widehat{\overline{\alpha}}_{\varepsilon}(c) = \frac{V((d_c^+(X_i))_{Y_i=0}, \varepsilon)}{n_0/n} :$$

a fractional knapsack. Move the closest cases first, and part of the last one if needed.

Backup: computation

- Logistic score: distances in closed form. For each threshold, sort the distances within each class once, and evaluate the dual at 0, at $1/\varepsilon$ and at the knots $1/d_i \in [0, 1/\varepsilon]$; each evaluation is $O(n)$.
- Posterior probabilities in closed form from two Beta CDFs, with no simulation; probabilities near 0 or 1 are compared on the log scale.
- Simulations: 1,001 thresholds, 5,000 validation cases, 500 doctors, 100 replications. Error rates by Gauss–Hermite quadrature with 1,024 nodes, which agree with 512 nodes to within 5×10^{-7}.
- Random forests: a search with K proposals (direction uniform on the sphere, length uniform on $[0, \rho]$). Searches of different sizes are nested, so their curves are comparable.

Backup: not a covariate shift

Covariate shift

$Q(Y \mid X) = P(Y \mid X)$; $Q(X)$ moves.

Importance weighting needs the deployment distribution, or a sample from it.

Here

$Q(Y)$ fixed; $Q(X \mid Y)$ moves; $Q(Y \mid X)$ may change.

The threshold is chosen before any deployment case is seen.

- A change in how a test result is recorded changes $Q(Y \mid X)$: the same recorded value now means something else.
- In the simulations, the shift changes $P(Y \mid X)$, while the prevalence and the doctors' error rates stay fixed.

Backup: Assumption 1 in full

Conditional on the independent training sample, the validation cases are independent draws from P_0, which satisfies the conditions of Proposition 1. Fix $\varepsilon > 0$ and a compact interval $\mathcal{C} = [c_l, c_u] \subset (0,1)$ of thresholds at which both decision regions are nonempty; $\Lambda = [0, 1/\varepsilon]$.

(i) $\mathcal{F}^{\pm} = \{\psi^{\pm}_{\lambda,c} : (\lambda, c) \in \Lambda \times \mathcal{C}\}$ are pointwise measurable with constant envelope M, and $\sup_H N(uM, \mathcal{F}^{\pm}, L_2(H)) \leq (K/u)^v$ for $0 < u < 1$;

(ii) on $\mathcal{C}$ the worst-case rates are continuously differentiable, with $0 < \underline{L}_\alpha \leq -\partial_c \overline{\alpha}_\varepsilon(c; P_0) \leq \overline{L}_\alpha < \infty$ and $|\partial_c \underline{\beta}_\varepsilon(c; P_0)| \leq L_\beta$; for the FPR interval $[\alpha_l, \alpha_u]$ there is $b_C > 0$ with $\overline{\alpha}_\varepsilon(c_l; P_0) > \alpha_u + b_C$ and $\overline{\alpha}_\varepsilon(c_u; P_0) < \alpha_l - b_C$.

Remark 2: for logistic scores, (i) follows from Lemma 3; (ii) needs the derivative bounds, for which Remark 3 gives primitive conditions.

Backup: Theorem 1, the four steps

1. *Dual objectives.* A maximal inequality over $\mathcal{F}^+ \cup \mathcal{F}^-$ (van der Vaart & Wellner 2011) gives $\sup_f |(\widehat{\mathbb{P}}_n - P_0)f| = O_p(n^{-1/2})$; replacing p_0 by $\widehat{p}$ costs another $O_p(n^{-1/2})$, uniformly on $\Lambda \times \mathcal{C}$.

2. *Rates.* $\sup_\lambda$ and $\inf_\lambda$ are 1-Lipschitz in the sup norm, so both worst-case rates inherit the rate.

3. *Inverse.* If $\delta_{\alpha,n} < b_C$, the smallest feasible threshold satisfies $c_{\alpha+\delta_{\alpha,n}} \le \widehat{c}_\alpha \le c_{\alpha-\delta_{\alpha,n}}$, and the inverse is Lipschitz with constant $1/\underline{L}_\alpha$.

4. *Curve.* $g_\varepsilon(\alpha - \delta_{\alpha,n}) - \delta_{\beta,n} \le \widehat{g}_\varepsilon(\alpha) \le g_\varepsilon(\alpha + \delta_{\alpha,n}) + \delta_{\beta,n}$, and the population curve is Lipschitz with constant $L_\beta/\underline{L}_\alpha$.

Numerical and grid errors add to these bounds; Theorem 1 assumes exact rates and a search over all thresholds.

Backup: the limit distribution of the estimated curve (Theorem 6)

- The worst-case ROC curve is a map from the moments $\mathcal{S}(P) = \left(PY,\ (Pf^{+}_{\lambda,c})_{\mathcal{K}},\ (Pf^{-}_{\lambda,c})_{\mathcal{K}}\right)$ to $\ell^{\infty}([\alpha_l, \alpha_u])$.
- Theorem 6: the map is Hadamard directionally differentiable when the sets of optimal multipliers $\Gamma_{\pm}(c)$ vary continuously in c; they need not be unique.
- Delta method for directionally differentiable maps (Fang & Santos 2019): $\sqrt{n}\left\{\widehat{g}_{\varepsilon} - g_{\varepsilon}(\cdot; P_0)\right\} \rightsquigarrow \dot{g}_{\mathbb{G}}$.
- Gaussian limit when the multipliers are unique, as for logistic scores under Remark 3 (Lemma 5). Otherwise the derivative can be nonlinear, and the limit alone does not justify the standard bootstrap.

Backup: Theorem 6, derivative formulas

Moments $\mathcal{S}(P)$ as before; directions $h = (h_p, h_+, h_-)$ with continuous $h_\pm$ on $\mathcal{K} = \Lambda \times \mathcal{C}$.

$$\dot{B}_h(c) = \frac{\max_{\lambda \in \Gamma_+(c)} h_+(\lambda, c) - B_0(c)\, h_p}{p_0}, \qquad \dot{A}_h(c) = \frac{\min_{\lambda \in \Gamma_-(c)} h_-(\lambda, c) + A_0(c)\, h_p}{1 - p_0},$$

$$\dot{c}_h(\alpha) = -\frac{\dot{A}_h(c_\alpha)}{A_0'(c_\alpha)}, \qquad \dot{g}_h(\alpha) = \dot{B}_h(c_\alpha) - \frac{B_0'(c_\alpha)}{A_0'(c_\alpha)}\, \dot{A}_h(c_\alpha).$$

- Needs $(\lambda, c) \mapsto f^{\pm}_{\lambda,c}$ continuous into $L_2(P_0)$ and $c \mapsto \Gamma_\pm(c)$ continuous in the Hausdorff distance; Lemma 4 expands an optimized value uniformly in c.
- Remark 3 and Lemma 5: for logistic scores with positive continuous index densities and budgets too small to move a whole side of the threshold, the multipliers are unique, $\lambda_+(c) = 1/\omega_1(a(c))$ and $\lambda_-(c) = 1/\omega_0(a(c))$, and the derivative is linear.

Backup: Lemma 6, posterior concentration

- Dirichlet posterior $\Rightarrow$ independent Beta posteriors, via four independent Gamma variables (the two ratios use disjoint pairs).
- Chernoff bound for a $\mathrm{Beta}(a, b)$ variable around its mean μ: $\mathbb{P}(|Z - \mu| > u) \le 2\exp\{-2(a+b)u^2\}$, using $\mathrm{kl}(\mu, x) \ge 2(x-\mu)^2$.
- Hoeffding for the empirical rates, the prior shrinkage bound $\|\mu_j - \tilde{\theta}_j\|_\infty \le a_0/(m_j + 2a_0)$, and a binomial bound on the smaller class count m_j.
- Result: $\mathbb{E}[T_j(t)] \le C_j \exp(-\zeta_j n_j t^2)$ with, for example,

$$C_j = \max\{8, e^{a_0/4}\} + 2, \qquad \zeta_j = \min\{p_*/16,\ p_*^2/2\},$$

where the prevalence lies in $[p_*, 1 - p_*]$.

Backup: Theorem 2, the exact bound

$$\mathbb{P}(\text{at least one wrong decision}) \;\le\; \mathbb{P}\big(r_n > \tfrac{\delta_0}{4}\big) + \frac{J\,C_0}{\bar{\tau}}\, e^{-\zeta_0\, n_{\min}\, t_0^2}$$

- $r_n = \sup_{\alpha\in[\alpha_l,\alpha_u]} \big|\widehat{g}_\varepsilon(\alpha) - g_\varepsilon(\alpha)\big|$: the largest error of the estimated worst-case ROC curve on the FPR range; $O_p(n^{-1/2})$ by Theorem 1, plus any grid error.
- δ_0: the gap. At the doctor's FPR, each doctor's TPR is at least δ_0 above or below g_ε (Assumption 2).
- J: the number of doctors; $n_{\min} = \min_j n_j$: the shortest history, in cases; $\bar{\tau} = \min\{\tau, 1-\tau\}$, which is 0.05 for the cutoff $\tau = 0.95$.
- $C_0 = \max_j C_j$ and $\zeta_0 = \min_j \zeta_j$, from Assumption 3: $\mathbb{E}[T_j(t)] \le C_j\, e^{-\zeta_j n_j t^2}$, where $T_j(t)$ is the posterior probability that doctor j's FPR or TPR is more than t from its true value.
- $0 < t_0 \le \min\{d_H/2,\ \delta_0/[4(1+L_g)]\}$: the radius of the neighborhood in the picture. Each doctor's FPR is at least d_H inside $[\alpha_l, \alpha_u]$, and g_ε is L_g-Lipschitz (Assumption 2).

back

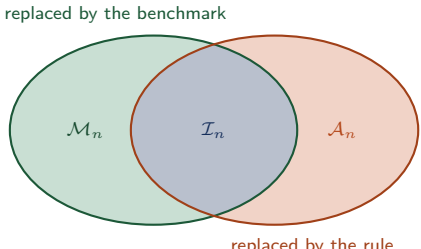
Backup: Theorem 2, proof sketch

Work on $\{r_n \le \delta_0/4\}$, and let K_j be the t_0-neighborhood (sup norm) of the true rates, $t_0 \le \min\{d_H/2,\ \delta_0/[4(1+L_g)]\}$.

- *Dominated doctor.* Take $\alpha^* = \alpha_{H,j} - t_0$ and the candidate c^* attaining $\widehat{g}_\varepsilon(\alpha^*)$ (Lemma 2). Then $\widehat{g}_\varepsilon(\alpha^*) \ge g_\varepsilon(\alpha_{H,j}) - L_g t_0 - r_n > \beta_{H,j} + t_0$, so the rectangle of c^* contains K_j, and the largest posterior probability is at least $1 - T_j(t_0)$.
- *Doctor not dominated.* If some rectangle met K_j, its candidate would be feasible at $\alpha_{H,j} + t_0$ with $\widehat{\underline{\beta}}_\varepsilon(c) \le \widehat{g}_\varepsilon(\alpha_{H,j} + t_0) \le g_\varepsilon(\alpha_{H,j}) + L_g t_0 + r_n < \beta_{H,j} - t_0$, a contradiction. So the largest posterior probability is at most $T_j(t_0)$.
- A wrong decision requires $T_j(t_0) \ge \bar{\tau} = \min\{\tau, 1-\tau\}$; Markov's inequality and a union bound give (3.7).

Backup: where regret comes from

$$
\begin{aligned}
&W(\mathcal{R}^*, c^*; Q) - W(\mathcal{R}_\varepsilon^{(n)}, \widehat{c}; Q) \\
&= \tfrac{1}{J} \sum_{j \in \mathcal{I}_n} \left[W_M(c_j^*; Q) - W_M(\widehat{c}_j; Q) \right] && \text{common} \\
&\quad + \tfrac{1}{J} \sum_{j \in \mathcal{M}_n} \left[W_M(c_j^*; Q) - W_{H,j} \right] && \text{missed} \\
&\quad + \tfrac{1}{J} \sum_{j \in \mathcal{A}_n} \left[W_{H,j} - W_M(\widehat{c}_j; Q) \right] && \text{extra}
\end{aligned}
$$



- The first sum is present even when the two sets of replaced doctors coincide, and its terms can have either sign.

Backup: Theorem 4(ii), the tolerances

With $t = \sqrt{\log\{2JC_0/(\gamma\bar{\tau})\}/(\zeta_0 n_{\min})} \le \min\{1, d_H\}$:

$$h_n = r_n + (1 + L_g)t, \qquad u_n = \phi e_{\beta,n} + \eta e_{\alpha,n} + (\phi + \eta)t, \qquad v_n = \phi r_n + \phi e_{\beta,n} + \eta e_{\alpha,n} + (\phi L_g + \eta)t.$$

- On $\{\max_j T_j(t) \le \bar{\tau}/2\}$ (probability $\ge 1 - \gamma$), every posterior median $a_{n,j}$ is within t of $\alpha_{H,j}$.
- Extra replacements: $0 \le \beta_{H,j} - g_Q(\alpha_{H,j}) \le h_n$; missed ones: $0 < g_Q(\alpha_{H,j}) - \beta_{H,j} \le \Delta_\varepsilon(Q) + h_n$. Assumption 4 bounds their shares by $C_M h_n^{\gamma_M}$ and $C_M\{\Delta_\varepsilon(Q) + h_n\}^{\gamma_M}$.
- Common replacements: $W_M(c_j^*; Q) - W_M(\widehat{c}_j; Q) \le \phi\Delta_\varepsilon(Q) + v_n + \phi\{\widehat{g}_\varepsilon(a_{n,j}) - \underline{\widehat{\beta}}_\varepsilon(\widehat{c}_j)\} + \eta\{\widehat{\overline{\alpha}}_\varepsilon(\widehat{c}_j) - a_{n,j}\}$; the last two terms give $\widehat{K}_n$ after taking positive parts.
- Proposition 4: if $r_n + e_{\alpha,n} + e_{\beta,n} = O_p(\sqrt{\log n/n})$, what remains after subtracting $\widehat{K}_n + \phi\widehat{\pi}_n\Delta_\varepsilon(Q)$ is $O_p(\bar{r}_n + \{\Delta_\varepsilon(Q) + \bar{r}_n\}^{1+\gamma_M})$, with $\bar{r}_n = \sqrt{\log n/n} + n_{\min}^{-1/2}$.

Backup: what the margin condition says

Assumption 4

There are $C_M, \gamma_M > 0$ such that, for every $Q \in \mathcal{B}_\varepsilon(P_0)$ and $t > 0$,
$\frac{1}{J}\sum_j \mathbb{I}(|g_Q(\alpha_{H,j}) - \beta_{H,j}| \le t) \le C_M t^{\gamma_M}$.

- Because J is fixed, a doctor at distance d from g_Q contributes $1/J$ at $t = d$, so the condition requires $\inf_Q \min_j |g_Q(\alpha_{H,j}) - \beta_{H,j}| \ge (JC_M)^{-1/\gamma_M} > 0$.
- It limits how many doctors lie near the algorithm's ROC curve, for a fixed group of doctors: the counterpart of the margin conditions of Mammen & Tsybakov (1999) and Audibert & Tsybakov (2007), with doctors in place of cases.
- Corollary 1 does not use it; it only bounds the missed and extra replacements.
- In Example 2 it holds with $\gamma_M = 1$ and $C_M = (0.6 - 2s)^{-1}$.

Backup: Example 2, why the selected threshold goes to $u = 0.5$

- Posterior probability at nominal rate u: $q_N(u) = F_{0,N}(1 - u - s)\, F_{1,N}(u - s)$, where $Z_{0,N} = 1 - \alpha_H$ and $Z_{1,N} = \beta_H$ are independent Beta variables with parameters $(a_0 + \mathrm{TN}_N, a_0 + \mathrm{FP}_N)$ and $(a_0 + \mathrm{TP}_N, a_0 + \mathrm{FN}_N)$.
- Beta tails: for $\mathrm{Beta}(a_N, b_N)$ with $a_N/N \to a$, $b_N/N \to b$ and $x > \theta = a/(a+b)$, $-\frac{1}{N} \log \mathbb{P}(Z_N > x) \to (a+b)\, \mathrm{kl}(\theta, x)$.
- Both tails have $(a, b) = (0.1, 0.4)$, so $\theta = 0.2$, by the symmetry of the doctor's errors ($\alpha_H = 0.8$, $\beta_H = 0.2$, prevalence $1/2$). At rate u they are evaluated at $1 - u - s$ and $u - s$: raising u fattens one tail and thins the other, so the chance of failing the test, $1 - q_N(u)$, vanishes fastest where the two rates balance, at $u = 0.5$.
- The welfare weights never enter. The posterior median of the FPR tends to 0.8, so $\widehat{K}_n \to 0.3 - s$.

Backup: Example 1, no threshold dominates under the average budget

	average budget 0.1	
threshold $x > t$	worst FPR	worst TPR
$t = 4.0$	0.747	0.665
$t = 4.5$	0.690	0.541
$t = 5.0$	0.600	0.400
$t = 5.5$	0.459	0.310
$t = 6.0$	0.335	0.253
$t = 6.5$	0.265	0.200
$t = 7.0$	0.224	0.150

- The doctor has rates $(0.35,\ 0.65)$.
- A worst-case FPR of at most 0.35 needs $t \geq 5.92$, where the worst-case TPR is at most 0.26.
- A worst-case TPR of at least 0.65 needs t near 4, where the worst-case FPR is about 0.75.
- Under the pointwise budget 0.1, every $t \in [4.8,\ 5.2)$ dominates the doctor.

Computed from the twenty cases of Example 1; not reported in the paper.

Backup: Theorem 5(iii), a counterexample at a jump level

- One-dimensional tree score $m(x) = 1/4$ for $x \le 0$, $m(x) = 3/4$ for $x > 0$; $\rho = 1/4$; $X \in \{-1, 1\}$ with $P_0(X = 1 \mid Y = 0) = a$ and $P_0(X = 1 \mid Y = 1) = b$.
- No ball of radius $1/4$ around a support point crosses the split, so $\underline{m}_\rho = \overline{m}_\rho = m$, and the attainable pairs are $(1, 1)$, (a, b), $(0, 0)$:

$$\tilde{g}_\rho(\alpha; P_0) = 0 \ (\alpha < a), \qquad b \ (a \le \alpha < 1), \qquad 1 \ (\alpha = 1).$$

- If $a = 1/2$, the empirical $\widehat{a}$ exceeds a with probability tending to $1/2$, and then the estimated curve is 0 at $\alpha = a$, while $\tilde{g}_\rho(a) = b$: not even pointwise consistent there.
- Continuous analogue: F_0 strictly increasing with $f_0(1/2) = 0$; an error of size t in the CDF moves the inverse by the *cube root* of t. The positive-density condition rules this out.

Backup: replacement using confidence bounds (Appendix A.4)

- Suppose upper bounds $U_{\alpha,n}(c) \geq \widehat{\overline{\alpha}}_{\varepsilon}(c)$ and lower bounds $L_{\beta,n}(c) \leq \widehat{\underline{\beta}}_{\varepsilon}(c)$ cover the worst-case rates at every candidate with probability $1-\gamma$.
- Use them in place of the estimated rates: $\widehat{q}_j^{\,L}(c) = \Pi_j\{\alpha_H \geq U_{\alpha,n}(c),\ \beta_H \leq L_{\beta,n}(c) \mid \mathcal{D}_j\}$; replace if $\max_c \widehat{q}_j^{\,L}(c) \geq \tau$.
- Proposition 5: (i) it replaces a subset of the doctors the implemented rule replaces; (ii) when the bounds cover, every replaced doctor is dominated at every Q in the set with posterior probability at least τ; (iii) it is consistent if the bounds shrink.
- Caveat: $1-\gamma$ is a frequentist coverage and τ a posterior cutoff; neither is an unconditional probability of dominance. The paper does not construct the bounds.

Backup: alternative welfare weights (Table B.1)

Robust ($\varepsilon \in \{0.05, 0.1, 0.2, 0.3\}$, nobody replaced) minus nominal, paired; the decisions do not depend on (ϕ, η).

δ	$(\phi,\eta)=(1,1)$	$(2,1)$	$(5,1)$
0	-0.0507 (0.0010)	-0.0826 (0.0017)	-0.1781 (0.0039)
0.025	-0.0445 (0.0009)	-0.0727 (0.0015)	-0.1575 (0.0034)
0.05	-0.0382 (0.0008)	-0.0628 (0.0013)	-0.1368 (0.0030)
0.1	-0.0254 (0.0005)	-0.0428 (0.0009)	-0.0951 (0.0023)
0.3	$+0.0275$ (0.0006)	$+0.0392$ (0.0010)	$+0.0745$ (0.0022)
0.5	$+0.0822$ (0.0017)	$+0.1228$ (0.0027)	$+0.2446$ (0.0055)
1	$+0.2168$ (0.0045)	$+0.3226$ (0.0068)	$+0.6398$ (0.0135)

The same signs under all three pairs of weights.

Backup: smaller average budgets (Table B.2)

Gain over retention, $(\phi, \eta) = (2, 1)$, means over 100 replications.

ε	replaced	$\delta = 0$	$\delta = \varepsilon$	$\delta = 0.05$	$\delta = 0.5$	$\delta = 1$
0	41.51%	0.08257	0.08257	0.06284	−0.12279	−0.32258
0.0001	22.79%	0.04628	0.04626	0.03532	−0.06860	−0.18150
0.0003	12.98%	0.02666	0.02662	0.02039	−0.03940	−0.10477
0.001	3.18%	0.00659	0.00656	0.00505	−0.00983	−0.02627
0.003	0.14%	0.00028	0.00028	0.00022	−0.00043	−0.00115
0.01, 0.05	0	0	0	0	0	0

At $\varepsilon = 0.003$, some doctor is replaced in 44 replications. Column $\delta = \varepsilon$: each rule at the edge of its own set (supplementary runs); it is within 0.00004 of $\delta = 0$, since these shifts are tiny.

Backup: harmful replacements under pointwise budgets (Table B.4)

Share of harmful replacements among replaced doctors, in percent; $(\phi, \eta) = (2, 1)$; †: inside the set ($\delta \leq \rho$).

	$\rho \backslash \delta$	0	0.025	0.05	0.1	0.3	0.5	1
$\widehat{a}^q$	0	0.17†	0.31	0.70	2.82	99.94	100	100
	0.05	0.24†	0.40†	0.74†	2.99	99.98	100	100
	0.1	0.13†	0.25†	0.65†	4.04†	100	100	100
	0.2	0.00†	0.00†	3.37†	11.52†	100	100	100
$\widehat{a}^w$	0	0.03†	0.06	0.17	0.86	98.71	100	100
	0.05	0.10†	0.12†	0.31†	1.11	99.66	100	100
	0.1	0.04†	0.07†	0.16†	1.60†	99.92	100	100
	0.2	0.00†	0.00†	2.13†	8.69†	100	100	100

Means over 100 replications for $\rho \leq 0.1$ and over the 47 with replacements for $\rho = 0.2$. The share counts harmful replacements, not the size of the losses: at $(\rho, \delta) = (0.2, 0.1)$ they are 0.016% of all doctors.

Backup: the nominal rule's regret at $\delta = 0$

- Regret relative to the matched-FPR benchmark: 0.17182 on average.
- Common replacements, which use different thresholds: 0.02857; missed replacements: 0.14325.
- No extra replacements: in all 100 replications every doctor lies below the algorithm's ROC curve under P_0 (the smallest gap is 0.048), so the benchmark replaces everyone.
- The threshold term $\widehat{K}_n$ averages 0.0327 (s.e. 0.0008).
- The robust rule with $\varepsilon = 0.05$ replaces nobody: its regret of 0.1864 at $\delta = 0.05$ comes entirely from missed replacements.

Backup: income data by budget and search size (Table C.1)

A. By budget (LR: average, exact; RF: pointwise, $K = 200$)

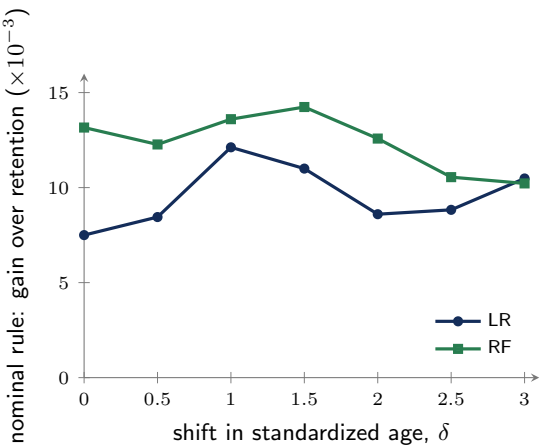
	area		largest post. prob.	
budget	LR	RF	LR	RF
0	0.8526	0.8790	0.9850	0.9907
0.1	0.2631	0.7386	0.0374	0.8954
0.2	0.0851	0.6337	1.03×10^{-4}	0.7032
0.3	0.0155	0.5230	2.79×10^{-9}	0.4298
0.5	0.0000	0.3937	0	0.1522

B. RF at $\rho = 0.5$, by number of proposals

K	area	largest post. prob.	replaced
20	0.5816	0.5834	0
50	0.5004	0.3855	0
100	0.4433	0.2485	0
200	0.3937	0.1522	0
500	0.3356	0.0786	0
1000	0.2968	0.0454	0

- A larger search lowers the RF curve, as it must. Exact extrema would also replace nobody at any positive budget.

Backup: shifting only the algorithm's data (age), further California cases



- 5,000 further California cases; labels and decisions fixed, so the decision makers' error rates do not change.
- δ is the distance $\mathcal{W}_\kappa$ between the shifted and unshifted cases, not the distance from P_0.
- Both nominal rules beat retention at every shift; the rules with budget 0.5 retain everyone.

Backup: critical budgets in the simulations

A larger budget replaces a subset of the doctors, so the share of the nominal rule's replacements that are still replaced at a budget is the share whose critical budget exceeds it. Pooled over the 100 replications:

average budget ε	0.0001	0.0003	0.001	0.003	0.01
still replaced	54.9%	31.3%	7.7%	0.3%	0
pointwise budget ρ	0.05	0.1	0.2	0.3	
still replaced	45.0%	13.5%	0.4%	0	

- Average budgets: the median critical budget lies between 0.0001 and 0.0003; almost none exceeds 0.003.
- Pointwise budgets: the median is below 0.05; 13.5% exceed 0.1.
- Income data, LR, about 30 cases: of the 5 replaced, 1 is still replaced at $\varepsilon = 0.001$ and none at 0.003.

Computed from the shares replaced in Tables B.2 and 2 (nominal 41.51%).

Backup: why this set, and how large a budget?

- **Why Wasserstein, not f-divergence?** A change in recording moves the data to new values; the f-divergence sets used in robust optimization, such as Kullback–Leibler or χ^2 balls, only reweight the values already seen.
- **Why label-preserving?** The comparison is between error rates, which condition on the label, and moving the data never changes whether a patient is sick.
- **Why not covariate shift?** A recording change alters $P(Y \mid X)$; covariate shift rules that out, and importance weighting needs the deployment distribution.
- **How large a budget?** No universal answer: report a grid and each doctor's critical budget; translate an average budget into rates with the $\sqrt{\varepsilon}$ scale; use a pointwise budget when a per-case bound is known.
- **Why a set rather than a prior?** All past cases come from the training distribution, so a prior on the shift would never be updated, and the decision would rest on the prior.

Backup: if one did put a prior on the set

Not in the paper; it follows from the definitions, at the population worst-case rates

- Let μ be any prior on $\mathcal{B}_\varepsilon(P_0)$, independent of the doctor's error rates. Past cases do not update it.
- For a fixed threshold c, every Q in the set has $\mathrm{FPR}_Q(c) \le \overline{\alpha}_\varepsilon(c)$ and $\mathrm{TPR}_Q(c) \ge \underline{\beta}_\varepsilon(c)$, so

$$\int \Pi_j\big(\alpha_H \ge \mathrm{FPR}_Q(c),\ \beta_H \le \mathrm{TPR}_Q(c) \mid \mathcal{D}_j\big)\,\mu(\mathrm{d}Q) \ \ge\ q_j^{\varepsilon}(c).$$

- The rule's posterior probability is a lower bound on the posterior probability of dominance under every such prior: a doctor replaced by the rule would be replaced under any prior on the set.

Backup: what if the doctors' error rates also change?

- Under (2.4), the posterior from the doctor's history is also a posterior for the error rates at deployment.
- Dropping it would require specifying each doctor's decision process in every environment: a model of how doctors respond to the new environment, not just a restriction on (X, Y).
- (2.4) restricts which environments are compared. The set restricts only the distribution of (X, Y), so it says nothing about the doctors.
- In the income data, (2.4) does not hold: the decision makers' decisions depend on the data seen by the algorithm.

Possible extension if asked (not in the paper): let the doctor's rates move within a box of known half-widths around the past rates; the inequalities at a fixed threshold would carry the half-widths as extra slack.

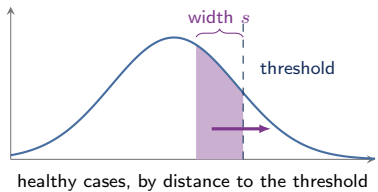
Backup: pointwise rules under the joint map, by size of the shift

Map $x + (1-2y)\, a\, w/\|w\|$: $a = \rho$ is the worst case in the set of the rule with budget ρ. Means over 100 replications, $(\phi, \eta) = (2, 1)$, threshold $\widehat{c}^{\,q}$; harmful shares over the replications with replacements (all 100, but 47 at $\rho = 0.2$).

		gain over retention			harmful among replaced	
ρ	replaced	$a = 0$	$a = \rho/2$	$a = \rho$	$a = \rho/2$	$a = \rho$
0.05	18.68%	0.0364	0.0312	0.0259	0.45%	0.97%
0.1	5.61%	0.0106	0.0074	0.0042	1.10%	7.01%
0.2	0.15%	0.00025	0.00008	−0.00009	13.65%	94.68%
0.3	0	0	0	0	–	–

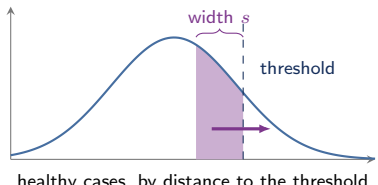
- Nominal rule, same maps: 0.0593 at $a = 0.05$, 0.0356 at 0.1, -0.0126 at 0.2, -0.0617 at 0.3.
- Welfare choice at $a = \rho$: 0.0300, 0.0052, -0.00008, 0; harmful 0.40%, 2.71%, 88.30%.
- $\rho = 0.2$: 74 replacements in 50,000 doctor-replications, none harmful at $a = 0$; s.e. of the loss 0.000017.

Backup: why small average budgets bite, $\sqrt{\varepsilon}$ versus ρ



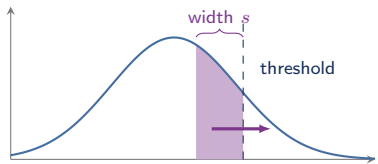
- Cheapest way to raise the FPR: push the strip of width s just below the threshold across it.

Backup: why small average budgets bite, $\sqrt{\varepsilon}$ versus ρ



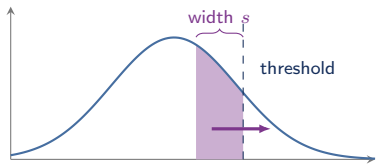
- Cheapest way to raise the FPR: push the strip of width s just below the threshold across it.
- The strip holds a share $f\,s$ of the healthy cases (f: their density at the threshold).

Backup: why small average budgets bite, $\sqrt{\varepsilon}$ versus ρ



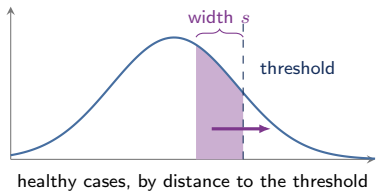
- Cheapest way to raise the FPR: push the strip of width s just below the threshold across it.
- The strip holds a share $f\,s$ of the healthy cases (f: their density at the threshold).
- Each moves $s/2$ on average: cost $\approx$ share $\times$ distance $= fs^2/2$.

Backup: why small average budgets bite, $\sqrt{\varepsilon}$ versus ρ

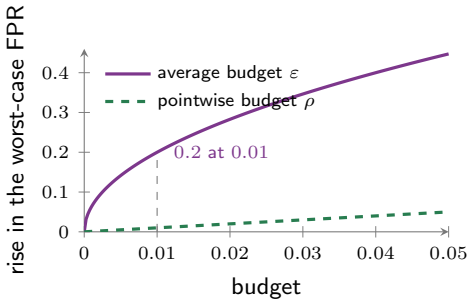


- Cheapest way to raise the FPR: push the strip of width s just below the threshold across it.
- The strip holds a share $f\,s$ of the healthy cases (f: their density at the threshold).
- Each moves $s/2$ on average: cost $\approx$ share $\times$ distance $= fs^2/2$.
- Spend the budget: $s \propto \sqrt{\varepsilon}$, so the FPR rises by $fs \propto \sqrt{\varepsilon}$.

Backup: why small average budgets bite, $\sqrt{\varepsilon}$ versus ρ

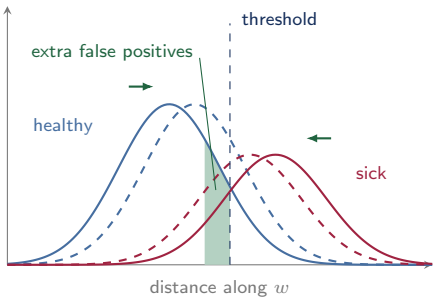


- Cheapest way to raise the FPR: push the strip of width s just below the threshold across it.
- The strip holds a share $f\,s$ of the healthy cases (f: their density at the threshold).
- Each moves $s/2$ on average: cost $\approx$ share $\times$ distance $= fs^2/2$.
- Spend the budget: $s \propto \sqrt{\varepsilon}$, so the FPR rises by $fs \propto \sqrt{\varepsilon}$.



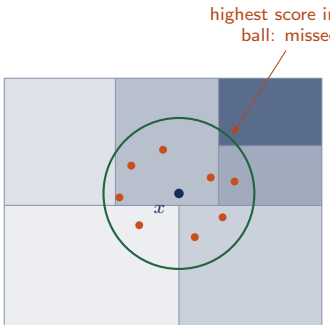
- Pointwise budget: only the cases within ρ can cross, so the FPR rises by $f\rho$.
- Example 2 (X uniform in each class, half the cases sick): average budget $0.01 \to 0.2$; pointwise budget $0.01 \to 0.01$.

Logistic scores: the two classes move toward each other



- For a logistic score,
  $\underline{m}_\rho(x) = \sigma\{z(x) - \rho\|w\|\}$ and
  $\overline{m}_\rho(x) = \sigma\{z(x) + \rho\|w\|\}$: each case moves a distance ρ along w.
- So the worst-case rates are the nominal ones with the threshold moved by ρ:
  $\tilde{\alpha}_\rho(c) = 1 - H_0(a_c - \rho)$,
  $\tilde{\beta}_\rho(c) = 1 - H_1(a_c + \rho)$.
- One map attains both worst cases at every threshold: $x + (1 - 2y)\,\rho\, w/\|w\|$.

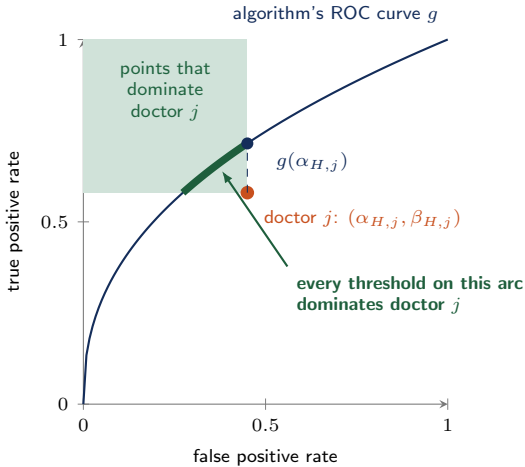
Tree-based scores: a search can only understate the worst case



- For random forests, the lowest and highest scores within ρ are found by a search over K random proposals in the ball.
- The search stays in the ball, so it can only understate the worst case. More data do not shrink this gap; a larger search can.
- **Theorem 5** and **Corollary 2** assume the lowest and highest scores are computed exactly: the curve is estimated at rate $n^{-1/2}$ (for trees, away from the FPR levels where it jumps), and the replacement and welfare results carry over.

Feng et al. (2025 MS): compare one threshold with one doctor

back



Some point dominates the doctor iff $\beta_{H,j} \leq g(\alpha_{H,j})$.

- Algorithm: score $m(X) \in [0,1]$; flag the case if $m(X) > c$. Varying the threshold c traces its ROC curve: $g(\alpha)$ is the highest TPR at an FPR of at most α.
- Doctor j: a single point, the false and true positive rates $(\alpha_{H,j}, \beta_{H,j})$.
- A point on the curve *dominates* the doctor if its FPR is no higher and its TPR no lower: at least as good whatever the costs of the two errors.
- The doctor's rates are estimated from past decisions. Replace the doctor if, for some point on the curve, the *posterior probability* that it dominates the doctor is at least τ, such as 0.95.