Robust Replacement of Human Decision Makers under Distribution Shift

Dingli Wang*

Abstract

This paper studies when a machine learning algorithm should replace a human decision maker under distribution shift. We infer each human's error rates from past decisions, assuming that they stay the same on future cases, while the data seen by the algorithm may shift within a Wasserstein ball. For each threshold, we compute the algorithm's worst false positive and true positive rates over the ball, which gives a worst-case ROC curve. We replace a human only if some point on this curve is, with high posterior probability, at least as good as the human in both rates. We show that the estimated curve converges uniformly and bound the welfare loss from replacement. We also compare the rule with an infeasible benchmark that knows the deployment distribution and each human's error rates and sets the algorithm's false positive rate equal to the human's. Relative to this benchmark, choosing the threshold with the highest posterior probability can leave a welfare loss that does not vanish with more data. In simulations, robust rules replace fewer humans than a rule that ignores shifts, giving up some gains without shifts but losing less after large shifts.

Keywords: Distributionally robust optimization, Wasserstein distance, ROC curve, human-algorithm comparison, distribution shift.
JEL Classification: C11, C12, C44, C45

*Department of Economics, University of California, Riverside, CA 92521, USA.
E-mail: dwang177@ucr.edu.

1 Introduction

Machine learning algorithms now perform as well as human experts in many prediction tasks (Esteva et al., 2017), and economists have used them to evaluate and improve human decisions in bail hearings and medical testing (Kleinberg et al., 2018; Mullainathan and Obermeyer, 2022) and have documented large differences in skill across decision makers (Chan et al., 2022). A natural next step is to ask when an algorithm should replace an individual human decision maker, such as a doctor. Comparisons between humans and algorithms usually evaluate the algorithm on held-out data from the same source as its training data. Once in use, however, an algorithm often faces cases from a different distribution, for example after a hospital changes how it records or extracts its data (Ötleş et al., 2021), and its performance can drop sharply (Zech et al., 2018; Finlayson et al., 2021; Wong et al., 2021). This paper studies the replacement decision when the distribution of the cases on which the algorithm will be used, which we call the deployment distribution, is not known in advance.

Our starting point is the replacement rule of Feng et al. (2025). They summarize each doctor by two error rates, the false positive rate (FPR) and the true positive rate (TPR), both estimated from the doctor's past decisions. They describe the algorithm by its receiver operating characteristic (ROC) curve, which shows the algorithm's FPR and TPR at each threshold. If a point on the curve has an FPR no higher and a TPR no lower than the doctor's, we say that the point dominates the doctor. At that threshold the algorithm is at least as good as the doctor, whatever the costs of the two errors. Because each doctor has only a limited history, they use a posterior for the doctor's error rates. They replace the doctor if, for some point on the curve, the posterior probability that this point dominates the doctor reaches a cutoff such as 0.95, and they then use the threshold at which this probability is highest. Their rule assumes that the deployment distribution is the same as the training distribution, the distribution of past cases.

We extend their rule to the case in which the deployment distribution is unknown. We assume that a doctor would have the same error rates on future cases as on past cases, while the data seen by the algorithm may change. For example, a hospital may change how it records a test result. Doctors examine patients in the same way as before, but the algorithm sees the test result recorded in the new way.

Our rule faces two kinds of uncertainty, and we treat them differently. A doctor's error rates are unknown but can be learned from the doctor's past decisions, so we use a posterior for them. The deployment distribution cannot be learned in this way. When we decide whether to replace a doctor, however many past cases we have, all of them come from the training distribution. The future cases that the algorithm will see have not arrived yet, and we impose no model that links the future shift to the past cases, so the past cases cannot tell us how the deployment distribution will differ from the training

distribution. Yet the deployment distribution determines how well the algorithm will perform after it replaces the doctor. We therefore specify a set of plausible deployment distributions, without assigning probabilities to them. We then require a high posterior probability that the algorithm, at one threshold chosen in advance, dominates the doctor under every distribution in the set.

We measure how far a distribution is from the training distribution by the Wasserstein distance, the smallest average amount by which the data seen by the algorithm must move to turn one distribution into the other. The set of plausible deployment distributions contains every distribution within a chosen distance of the training distribution, which we call the budget. Moving the data seen by the algorithm never changes whether a patient is sick, so every distribution in the set has the same share of sick patients as the training distribution. For each threshold, we then ask how high the algorithm's FPR and how low its TPR could be in this set. To raise the FPR, the budget is best spent on the healthy patients closest to the threshold, and to lower the TPR, on the sick patients closest to it. At a given threshold, the algorithm dominates the doctor under every distribution in the set exactly when its worst-case FPR is no higher and its worst-case TPR no lower than the doctor's. Plotting the worst-case rates across thresholds gives a worst-case ROC curve. We replace a doctor if, for some point on this curve, the posterior probability that the point dominates the doctor reaches the cutoff. Several thresholds may pass this test. We then use the one with the highest posterior probability of dominance, which does not require the costs of the two errors.

To apply this rule, we estimate the worst-case rates from data. We compute them on a validation sample, which consists of past cases that were set aside when the algorithm was trained. We show that these estimates, and the estimated worst-case ROC curve, converge to their true values at the usual root-n rate, uniformly over a range of thresholds. This requires two conditions. The first limits the complexity of the algorithm's score, in the usual sense of empirical process theory, and holds for logistic scores. The second requires the worst-case FPR to fall steadily as the threshold rises, so that each FPR corresponds to a single threshold. Suppose also that no doctor lies exactly on the true worst-case ROC curve. Then, as the validation sample and the doctors' histories grow, the probability that the rule replaces exactly the doctors below the curve approaches one. The closer a doctor is to the curve, the more data this takes. Under further conditions, we also derive the limit distribution of the estimated worst-case ROC curve.

Dominance ensures that the algorithm is at least as good as the doctor whatever the costs of the two errors. To measure how much replacement gains or loses, we need these costs, and we use welfare, the value of true positives minus the cost of false positives. Because the worst-case rates and the doctors' error rates are estimated, the selected threshold dominates a replaced doctor only up to estimation error, so replacement can still lower welfare. We show that, under every plausible deployment distribution, the

possible loss relative to retaining all doctors is bounded by these estimation errors and shrinks as the validation sample and the doctors' histories grow.

To see what our rule gives up, we also compare it with a benchmark that knows the deployment distribution and each doctor's true error rates and so cannot be implemented. Because it faces no uncertainty, the benchmark replaces exactly the doctors whom some point on the algorithm's ROC curve under the deployment distribution dominates. For each of them, it uses the threshold at which the algorithm has the same FPR as the doctor. The algorithm then makes no more false positives than the doctor and finds as many sick patients as it can. We call the welfare that our rule gives up relative to the benchmark its regret. We bound the regret by three terms. The first is how far the worst-case ROC curve lies below the algorithm's ROC curve under the deployment distribution, which is the price of guarding against shifts. The second comes from the estimation errors, and the third comes from the choice of threshold. The bound uses a margin condition at the deployment distribution: few doctors lie close to the algorithm's ROC curve under that distribution, so small errors change few replacement decisions.

The third term need not vanish. The threshold with the highest posterior probability of dominance is the one at which dominance is most certain, but it need not have the highest welfare. We give an example in which all estimation errors vanish: the worst-case rates are known, and the doctor's history grows without bound. Even so, the rule's threshold does not approach the benchmark's, and the third term does not go to zero. If the costs of the two errors can be specified, we can instead choose, among the thresholds that pass the posterior test, the one with the highest welfare lower bound computed from the worst-case rates. This welfare choice replaces the same doctors and changes only the threshold.

All the results above use a set that limits only the average movement of the data seen by the algorithm. We call this limit an average budget. We also study a pointwise budget, which limits how far each case can move. A pointwise budget fits shifts that move each case by at most a known distance, for example after a change in measurement equipment. An average budget of the same size covers all of these shifts, and also shifts that move a few cases much farther. The replacement and welfare results above carry over to the pointwise budget. This budget needs only the lowest and highest scores near each case, so it works for tree-based scores such as random forests. For these scores we find the lowest and highest scores by a search, which can only understate the worst case. For logistic scores, the worst-case rates under both budgets can be computed exactly, so we can compare how much each budget raises the worst-case FPR at a fixed threshold. For a small budget, a pointwise budget raises it above the FPR without a shift by an amount proportional to the budget. An average budget of the same size raises it by an amount proportional to the square root of the budget, which is much larger. The reason is that an average budget can spend all of its movement on the few cases closest to the threshold.

The worst-case TPR falls in the same way.

In the simulations, the shift moves the data seen by the algorithm for every case by the same distance, in opposite directions for sick and healthy patients, while the doctors' error rates stay the same. We compare our rule with the nominal rule, which sets the budget to zero and is the rule of Feng et al. (2025). The nominal rule raises mean welfare after small shifts, but after large shifts almost all of its replacements lower welfare and it does worse than retaining all doctors. With the prespecified average budgets our rule retains all doctors, giving up the nominal rule's gains after small shifts and avoiding its losses after large ones. With much smaller average budgets, and with pointwise budgets, it replaces some doctors, gains less than the nominal rule after small shifts and loses less after large ones; with pointwise budgets of 0.05 and 0.1 it also raises mean welfare under the worst case in its set. When the costs of the two errors are specified, the welfare choice has higher mean welfare than the selected threshold, at the cost of a lower posterior probability of dominance.

We also apply the rules to income data from US states, with synthetic decision makers. With histories of about 30 cases each, the nominal rule replaces few decision makers and changes welfare only slightly, and our rule replaces no one with average budgets of 0.003 or more. With histories of 300 cases, the nominal rule replaces many more decision makers, but these budgets still replace no one, so the short histories explain only part of the result.

Our worst-case rates build on optimal-transport risk bounds (Blanchet and Murthy, 2019; Gao and Kleywegt, 2023; Mohajerin Esfahani and Kuhn, 2018). Shafieezadeh-Abadeh et al. (2015, 2019) derive worst- and best-case misclassification bounds for a fixed linear classifier. For logistic scores, the algorithm's decision boundary is also a hyperplane, and our worst-case rates use the same distances to it, computed within each class. We combine these bounds with each doctor's posterior to decide whom to replace and which threshold to use.

This paper is related to policy choice under ambiguity (Manski, 2011). Adjaho and Christensen (2022) learn treatment policies that are robust to shifts in the joint distribution of potential outcomes and characteristics within a Wasserstein neighborhood of the experimental distribution. Kido (2022) studies Wasserstein-robust treatment-policy learning using experimental data from a source population and covariate data from the target population. Mo et al. (2021) study individualized treatment rules under covariate shifts using uncertainty sets that bound an L^k norm of the covariate density ratio. We study a different decision: whether a machine threshold fixed before deployment has sufficient posterior support to dominate a particular human in both error rates. The main replacement criterion does not require relative error costs; welfare-based threshold choice is considered separately.

More broadly, Kuhn et al. (2019) review distributionally robust optimization over Wasserstein balls, and Duchi and Namkoong (2021) study uniform performance over

neighborhoods defined by f-divergences. A different approach to distribution shift, importance weighting, reweights the training cases to match the deployment distribution and therefore needs this distribution or a sample from it (Sugiyama and Kawanabe, 2012). In our setting, the threshold is chosen before any case from the deployment distribution is observed, so we guard against a set of distributions instead. Our margin condition is the counterpart of the margin conditions for classifiers (Mammen and Tsybakov, 1999; Audibert and Tsybakov, 2007). Those conditions limit how many cases lie near the decision boundary. Ours limits how many doctors lie near the algorithm's ROC curve, for a fixed group of doctors. The analysis is also related to work on binary decision problems and ROC inference (Elliott and Lieli, 2013; Feng et al., 2026).

Section 2 presents the decision rule, and Section 3 gives the statistical and welfare results. Section 4 studies pointwise budgets. Sections 5 and 6 report the simulations and the income-data illustration, and Section 7 concludes. Appendix A contains proofs and additional inference results, and Appendices B and C give numerical details.

2 The replacement decision under distribution shift

2.1 Setup

There are J human decision makers, indexed by $j = 1, \ldots, J$, whom we call doctors, as in the introduction. Each case has data $X \in \mathcal{X} \subseteq \mathbb{R}^d$ seen by the algorithm and a label $Y \in \{0, 1\}$, which equals one if the patient is sick. Let P_0 denote the distribution of (X, Y) in past cases, which we call the training distribution. The algorithm's score $m(x)$ is fitted on a training sample drawn from P_0, and we treat m as fixed. A separate validation sample, drawn from P_0, is used to estimate the algorithm's error rates. For each doctor j, we also observe a history $\mathcal{D}_j$ of past cases drawn from P_0, which records for each case whether the patient was sick and what the doctor decided. On the basis of these data, a planner, such as a hospital, decides for each doctor whether to retain the doctor or to replace the doctor with the algorithm at a single threshold. The threshold is chosen before the deployment distribution is known and is not adjusted afterwards.

For case i assigned to doctor j, $\widehat{Y}_{i,j} \in \{0, 1\}$ is the doctor's decision and $\widehat{Y}_{i,M} \in \{0, 1\}$ is the algorithm's decision. The subscript H marks the doctor's rates, and the subscript M marks the algorithm's. Doctor j's false positive and true positive rates are

$$\alpha_{H,j} = \frac{\mathbb{E}[(1 - Y_i)\widehat{Y}_{i,j}]}{1 - p}, \qquad \beta_{H,j} = \frac{\mathbb{E}[Y_i \widehat{Y}_{i,j}]}{p}, \tag{2.1}$$

where $p = p_0 = P_0(Y = 1)$ is the prevalence, the share of sick patients, and the expectations in (2.1) are taken under the doctor's decision process in past cases. We write $\theta_{H,j} = (\alpha_{H,j}, \beta_{H,j})$.

The score $m(x)$ takes values in $[0,1]$, and the algorithm decides by the strict threshold rule $\widehat{Y}_{i,M} = \mathbb{I}(m(X_i) > c)$, where $\mathbb{I}(A)$ is the indicator of the event A. We define its ROC curve through an FPR constraint:

$$g(\alpha) = \sup_{c\in[0,1]} \left\{ \frac{\mathbb{E}[Y \cdot \mathbb{I}(m(X) > c)]}{p} \;\middle|\; \frac{\mathbb{E}[(1-Y) \cdot \mathbb{I}(m(X) > c)]}{1-p} \leq \alpha \right\}. \tag{2.2}$$

The two ratios in (2.2) are the algorithm's TPR and FPR at threshold c, so $g(\alpha)$ is the highest TPR that the algorithm can reach with a threshold whose FPR is at most α. When the score of healthy cases is continuously distributed, as for a logistic score with continuous data, every FPR level is reached by some threshold, and the constraint in (2.2) holds with equality.[1] When the score of healthy cases takes some values with positive probability, as with a random forest, the FPR jumps at these values, and some FPR levels are reached by no threshold. Under the inequality in (2.2), $g(\alpha)$ is then the TPR of the best threshold whose FPR does not exceed α. We do not fill these gaps by choosing at random between two thresholds, as some definitions of the ROC curve do, because the algorithm uses a single threshold. Because the algorithm will also be evaluated under distributions other than P_0, we write, for any distribution Q of (X, Y), $\mathrm{FPR}_Q(c) = Q(m(X) > c \mid Y = 0)$ and $\mathrm{TPR}_Q(c) = Q(m(X) > c \mid Y = 1)$, and we write $g_Q(\alpha)$, or $g(\alpha; Q)$, for the curve (2.2) under Q. Without a subscript, g denotes the curve under P_0.

Feng et al. (2025) replace doctor j if and only if

$$\widehat{q}_{\max,j} := \max_{\alpha\in[0,1]} \Pi_j\{\alpha_H \geq \alpha,\ \beta_H \leq g(\alpha) \mid \mathcal{D}_j\} \geq \tau, \tag{2.3}$$

where $\Pi_j\{\cdot \mid \mathcal{D}_j\}$ denotes the posterior probability given the history $\mathcal{D}_j$, obtained from a Dirichlet prior on the four cells of the doctor's confusion matrix (true and false positives and negatives; see Section 2.4). Under this posterior, $\theta_H = (\alpha_H, \beta_H)$ denotes the doctor's error rates as a random vector, and $\theta_{H,j}$ is their true value. The cutoff $\tau \in (0,1)$ is the required posterior probability of dominance. In words, the doctor is replaced if, for some point on the ROC curve, the posterior probability that this point dominates the doctor is at least τ.

Doctors' error rates under the deployment distribution. The deployment distribution describes the future cases, not how the doctors would decide on them. We call a distribution of (X, Y) together with a decision process for each doctor an environment. The set of plausible deployment distributions introduced below restricts only the distribution of (X, Y), so it says nothing about the doctors. We therefore assume that, in every

[1] If $m(x) = \mathbb{P}(Y = 1 \mid X = x)$ and the score has continuous distributions given Y, the curve (2.2) is the Neyman–Pearson ROC curve based on X: no classifier based on X has a higher TPR at the same FPR.

environment we consider, each doctor's error rates are the same as in past cases:

$$\mathbb{P}_Q(\widehat{Y}_j = 1 \mid Y = 0) = \alpha_{H,j}, \qquad \mathbb{P}_Q(\widehat{Y}_j = 1 \mid Y = 1) = \beta_{H,j}. \tag{2.4}$$

Here $\mathbb{P}_Q$ is the probability in the environment in which (X, Y) has distribution Q, including the doctor's decision process. Under (2.4), the posterior computed from the doctor's history is also a posterior for the doctor's error rates on future cases.

A simple sufficient condition is the following. Doctor j decides by a rule h_j applied to the information Ω_j that the doctor uses. This information can include what the algorithm does not see, such as what the doctor learns by examining the patient. The rule h_j and the distribution of Ω_j given Y are the same in every environment, while the distribution of X given Y may change. The doctor's error rates are then the means of $h_j(\Omega_j)$ given $Y = 0$ and given $Y = 1$, and these do not change. For example, a hospital may change how diagnoses are coded in its records, or a laboratory may recalibrate an assay and report new reference ranges. Doctors who examine patients, or who read results against the reported ranges, can keep their error rates, while the data seen by the algorithm no longer follow the distribution of the records it was trained on. Condition (2.4) can fail when doctors rely on the same shifted records without adjusting for the change.

The doctor's history gives a likelihood for the doctor's error rates, which we combine with the Dirichlet prior to form the posterior. For the data seen by the algorithm, the planner instead specifies a set of plausible deployment distributions, without attaching probabilities to them. The replacement rule of Section 2.3 combines the two: it requires a high posterior probability that, at one threshold, the algorithm dominates the doctor under every distribution in the set.

2.2 The worst-case ROC curve

We now define the set of plausible deployment distributions and the algorithm's worst-case rates over this set. The set is a ball around the training distribution P_0. In practice we use the empirical distribution $\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$ of the n validation cases in place of P_0, where δ_z denotes the point mass at z. We therefore state the definitions for a general center P, which is either P_0 or $\widehat{\mathbb{P}}_n$, and from now on the set means the ball around P.

Let $\mathcal{Z} := \mathcal{X} \times \{0, 1\}$ be the set of possible values of (X, Y). For two distributions P and Q on $\mathcal{Z}$ and a transport cost $\kappa : \mathcal{Z} \times \mathcal{Z} \to [0, \infty]$, define

$$\mathcal{W}_\kappa(P, Q) = \inf_{\pi \in \Pi(P,Q)} \int \kappa(z, z')\, d\pi(z, z'), \tag{2.5}$$

where $\Pi(P, Q)$ is the set of couplings with marginals P and Q. A coupling is a plan for moving the mass of P so that it becomes Q, and $\mathcal{W}_\kappa(P, Q)$ is the lowest cost of such a plan when moving a unit of mass from z to z' costs $\kappa(z, z')$.

We use a label-preserving transport cost:

$$\kappa\big((x,y),(x',y')\big) = \begin{cases} \|x - x'\|_2, & y = y', \\ +\infty, & y \neq y'. \end{cases} \tag{2.6}$$

Under this cost, moving a unit of mass within a class, $Y = 0$ (healthy) or $Y = 1$ (sick), costs the distance it moves, and mass cannot move to the other class: moving the data seen by the algorithm never changes whether a patient is sick. Write $\mathcal{P}(\mathcal{Z})$ for the set of all distributions of (X, Y). For $P \in \mathcal{P}(\mathcal{Z})$, the Wasserstein ball of radius ε around P is

$$\mathcal{B}_\varepsilon(P) = \{Q \in \mathcal{P}(\mathcal{Z}) : \mathcal{W}_\kappa(Q, P) \leq \varepsilon\}. \tag{2.7}$$

We call ε the budget: the larger the budget, the larger the shifts under which the algorithm must dominate the doctor. The set of plausible deployment distributions is $\mathcal{B}_\varepsilon(P_0)$. In the literature on distributionally robust optimization, such a set is called an ambiguity set (Kuhn et al., 2019).

Because no mass moves between the classes, every Q with $\mathcal{W}_\kappa(P, Q) < \infty$ has the same prevalence as P: $Q(Y = 1) = P(Y = 1)$. Let P^y and Q^y denote the conditional distributions of X given $Y = y$, and let W_1 be the Wasserstein distance of order one between distributions on $\mathcal{X}$, defined as in (2.5) with cost $\|x - x'\|_2$. With $p = P(Y = 1) \in (0, 1)$, $\mathcal{W}_\kappa$ separates by class:

$$\mathcal{W}_\kappa(P, Q) = pW_1(P^1, Q^1) + (1 - p)W_1(P^0, Q^0). \tag{2.8}$$

Hence a distribution in the set differs from P only in the distribution of X within each class. Distributions in the set need not satisfy $Q(Y \mid X) = P(Y \mid X)$, the condition that defines covariate shift (Shimodaira, 2000; Sugiyama et al., 2007).

Definition 1 (Worst-case ROC curve)**.** *Let P be a distribution on (X, Y) with $p := P(Y = 1) \in (0, 1)$. For each threshold $c \in [0, 1]$, define*

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

$$\overline{\alpha}_\varepsilon(c; P) := \sup_{Q \in \mathcal{B}_\varepsilon(P)} Q\big(m(X) > c \mid Y = 0\big).$$

The worst-case ROC curve is

$$g_\varepsilon(\alpha; P) := \sup\left\{\underline{\beta}_\varepsilon(c; P) : \overline{\alpha}_\varepsilon(c; P) \leq \alpha\right\}, \qquad \alpha \in [0, 1]. \tag{2.9}$$

We call $\underline{\beta}_\varepsilon(c; P)$ and $\overline{\alpha}_\varepsilon(c; P)$ the worst-case TPR and the worst-case FPR at threshold c, or jointly the worst-case rates. A threshold is feasible at level α if its worst-case

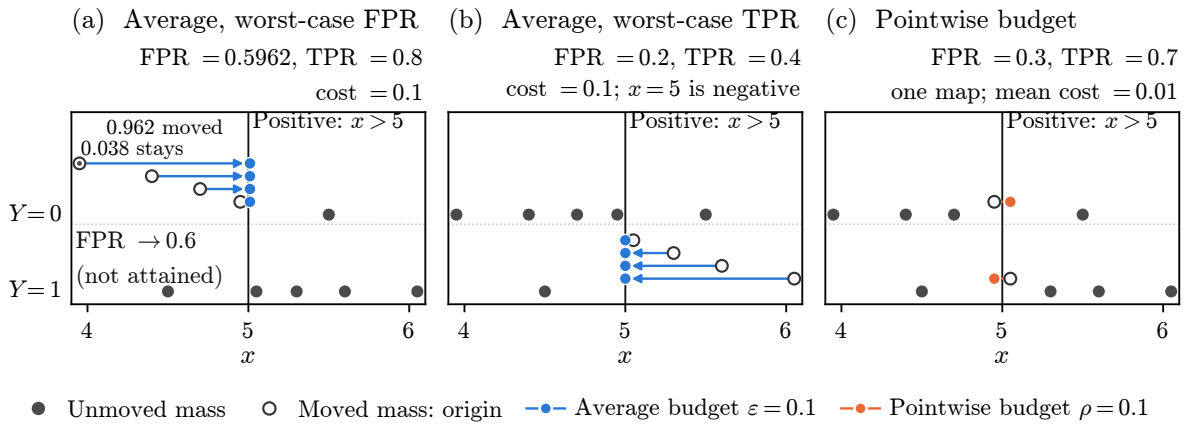


Figure 1: Average and pointwise budgets in Example 1.

Notes: The panels show the cases between $x = 3.9$ and $x = 6.1$; the other cases do not move. A case is classified as positive when $x > 5$, so a case at $x = 5$ is negative. (a) Worst-case FPR under the average budget $\varepsilon = 0.1$: the healthy cases at 4.4, 4.7 and 4.95 move to $5 + h$, and a fraction $f(h) = (1.05 - 3h)/(1.05 + h)$ of the case at 3.95 moves there, which uses the whole budget. With $h = 0.01$, as drawn, $f(h) \approx 0.962$ and the FPR is 0.5962; as $h \downarrow 0$, the FPR approaches 0.6, which is not attained. (b) Worst-case TPR under the same budget: the sick cases at 5.05, 5.3, 5.6 and 6.05 move to $x = 5$, and the TPR is 0.4. Panels (a) and (b) show two different distributions, each of which uses the whole budget. (c) Pointwise budget $\rho = 0.1$: moving the healthy case at 4.95 to 5.05 and the sick case at 5.05 to 4.95 gives the worst-case FPR of 0.3 and the worst-case TPR of 0.7 in one distribution. Vertical offsets within a lane only separate the arrows.

FPR is at most α, that is, if $\mathrm{FPR}_Q(c) \leq \alpha$ for every Q in the set. Then $g_\varepsilon(\alpha; P)$ is the highest TPR that such a threshold, chosen before deployment, can guarantee under every distribution in the set.

To see what the worst cases look like, fix a threshold. Moving a healthy case above the threshold raises the FPR, and the cheapest cases to move are those closest to it. To find the worst-case FPR, we therefore move healthy cases across the threshold, starting with the closest. Moving sick cases would use budget without changing the FPR, so we leave them in place. To find the worst-case TPR, we start again from P with the whole budget and move only sick cases, from just above the threshold onto it. The two worst cases therefore need not come from the same distribution in the set, since a single distribution must share the budget between the two classes, as (2.8) shows. Computing them separately is what the comparison with a doctor needs: at a given threshold and under (2.4), the algorithm dominates the doctor under every distribution in the set exactly when its worst-case FPR is at most the doctor's FPR and its worst-case TPR is at least the doctor's TPR. Example 1 and Figure 1 illustrate these points, and Section 2.4 describes the computation in general.

Example 1 (Worst-case rates in a small sample)**.** *Let P be the empirical distribution of*

$n = 20$ validation cases with a scalar X, ten healthy ($Y = 0$) and ten sick ($Y = 1$):

$$Y = 0: \ 0.5,\ 1,\ 2,\ 3,\ 3.95,\ 4.4,\ 4.7,\ 4.95,\ 5.5,\ 7;$$
$$Y = 1: \ 3,\ 4.5,\ 5.05,\ 5.3,\ 5.6,\ 6.05,\ 7,\ 8,\ 9,\ 9.5.$$

Let $m(x) = \{1 + \exp(5 - x)\}^{-1}$ and $c = 1/2$, so a case is classified as positive when $x > 5$. Two healthy and eight sick cases lie above 5, so under P the FPR is 0.2 and the TPR is 0.8.

Each case has mass $1/20$, so moving one case a distance d costs $d/20$, and any moves whose total cost is at most ε give a distribution in the set. Take $\varepsilon = 0.1$, so the distances moved can add up to at most 2. The four healthy cases closest to the threshold lie 0.05, 0.3, 0.6 and 1.05 below it, and these distances add up to 2. The budget is just enough to bring them to 5. Because a case exactly at 5 is still classified as negative, the FPR cannot reach 0.6. Moving the three closest cases just past 5 and almost all the mass of the fourth brings it arbitrarily close, so $\overline{\alpha}_\varepsilon(c; P) = 0.6$. Moving the sick cases at 5.05, 5.3, 5.6 and 6.05 to 5 gives $\underline{\beta}_\varepsilon(c; P) = 0.4$. Because the distances moved in the two classes must together stay within 2, no distribution in the set comes close to both rates at once. Panels (a) and (b) of Figure 1 illustrate the two separate worst-case calculations.

Suppose instead that each case may move by at most 0.1, as under the pointwise budget of Section 4.1. Only the healthy case at 4.95 and the sick case at 5.05 can then cross the threshold, so the worst-case FPR is 0.3 and the worst-case TPR is 0.7 (panel (c)). A doctor with rates $(0.35, 0.65)$ is dominated by this threshold under the pointwise budget. Under the average budget $\varepsilon = 0.1$ of (2.7), no threshold dominates the doctor, because the budget can be spent on the few cases closest to each threshold.

The next result gives both worst-case rates for a general P.

Proposition 1 (Exact dual representation of the worst-case rates at a fixed threshold)**.** *Let $\mathcal{X}$ be a closed subset of $\mathbb{R}^d$, let P have a finite first moment and $p \in (0, 1)$, and let m be continuous. Fix $\varepsilon > 0$ and a threshold c for which $S_c := \{x \in \mathcal{X} : m(x) > c\}$ and its complement are both nonempty. Define distances within $\mathcal{X}$ by*

$$d_c^-(x) := \inf\{\|x - z\|_2 : m(z) \le c\}, \qquad d_c^+(x) := \inf\{\|x - z\|_2 : m(z) > c\}.$$

Then

$$\underline{\beta}_\varepsilon(c; P) = \sup_{\lambda \ge 0} \left\{ \frac{1}{p} \mathbb{E}_P\big[Y \cdot \min\{1, \lambda d_c^-(X)\}\big] - \frac{\lambda}{p}\,\varepsilon \right\}, \tag{2.10}$$

and

$$\overline{\alpha}_\varepsilon(c; P) = \inf_{\lambda \ge 0} \left\{ \frac{\lambda}{1 - p}\,\varepsilon + \frac{1}{1 - p} \mathbb{E}_P\big[(1 - Y) \cdot (1 - \lambda d_c^+(X))_+\big] \right\}, \tag{2.11}$$

where $(u)_+ = \max\{u, 0\}$ for a real number u. Both optimizations can be restricted to $0 \le \lambda \le 1/\varepsilon$. The same interval applies when P is replaced by an empirical distribution

containing both classes. If S_c is empty, both rates are zero; if $S_c = \mathcal{X}$, both rates are one.

Here $d_c^-(x)$ is how far a case at x must move before the algorithm classifies it as negative, and $d_c^+(x)$ is how far it must move before the algorithm classifies it as positive. The number λ is the Lagrange multiplier on the budget, a price per unit of distance moved. When the whole budget is spent on the sick cases, they can move by ε/p on average, and the term $\lambda\varepsilon/p$ in (2.10) is the value of this movement at price λ. The term $\lambda\varepsilon/(1-p)$ in (2.11) plays the same role for the healthy cases. Moving a case across the threshold changes its contribution to the rate by one and costs λ times its distance, so at a given λ, moving a case pays exactly when its distance is below $1/\lambda$. The closest cases are thus moved first, as described before Example 1. Each worst-case rate is therefore found by a search over the single number λ instead of a search over all distributions in the set.

We prove Proposition 1 one class at a time, using the duality result of Blanchet and Murthy (2019) for the probability of a closed set. Because S_c is open, we also show that, when $\varepsilon > 0$, the supremum of the probability of S_c over the set equals that of its closure, although a worst-case distribution need not exist (Example 1).

At $\varepsilon = 0$ the set contains only P, so the worst-case rates are the rates under P, which we call the nominal rates:

$$\underline{\beta}_0(c; P) = P^1(S_c), \qquad \overline{\alpha}_0(c; P) = P^0(S_c).$$

We compute them from this display rather than by setting $\varepsilon = 0$ in (2.11), which gives the probability $P^0(\overline{S}_c)$ of the closure of S_c. A healthy case on the boundary of S_c is classified as negative, but its distance $d_c^+(x)$ to S_c is zero, so (2.11) counts it. The two values differ exactly when $P^0(\overline{S}_c \setminus S_c) > 0$, which can happen with an empirical distribution when a threshold equals the score of a healthy validation case. The formulas in Proposition 1 also exclude constant rules, which classify every case the same way, so that one decision region is empty and the distance to it is infinite. Their rates are 0 or 1 under every distribution in the set, as stated in Proposition 1, and we use these values directly.

2.3 The robust replacement decision

We use the worst-case ROC curve to compare the algorithm with a doctor.

Definition 2 (Robust dominance)**.** *A point $\theta_M = (\alpha_M, \beta_M)$ on the worst-case ROC curve $g_\varepsilon(\cdot)$ robustly dominates a doctor with error rates $\theta_H = (\alpha_H, \beta_H)$ if $\alpha_M \le \alpha_H$ and $\beta_M \ge \beta_H$.*

It follows that some point on the worst-case ROC curve robustly dominates the doctor if and only if $\beta_H \le g_\varepsilon(\alpha_H)$. For a fixed threshold c, the event

$$\overline{\alpha}_\varepsilon(c; P) \le \alpha_H, \qquad \underline{\beta}_\varepsilon(c; P) \ge \beta_H$$

means that, under (2.4), c dominates the doctor under every distribution in the set, as noted in Section 2.2. Under the posterior Π_j of Section 2.1, the probability of this event is

$$q_j^{\varepsilon}(c;P) := \Pi_j\Big(\alpha_H \geq \overline{\alpha}_{\varepsilon}(c;P),\ \beta_H \leq \underline{\beta}_{\varepsilon}(c;P) \mid \mathcal{D}_j\Big)\,. \tag{2.12}$$

In terms of the doctor's error rates, this event is the rectangle $[\overline{\alpha}_{\varepsilon}(c;P),1]\times[0,\underline{\beta}_{\varepsilon}(c;P)]$, and (2.12) is its posterior probability. The probability is computed for one threshold at a time, not for the whole region below the worst-case ROC curve. Figure 2 shows the rectangle for two doctors in the simulations of Section 5.

Definition 3 (Robust Bayesian replacement)**.** *Given $g_{\varepsilon}(\cdot;P)$ and the posterior Π_j, define*

$$\widehat{q}^{\varepsilon}_{\max,j} := \max_{\alpha\in[0,1]} \Pi_j(\alpha_H \geq \alpha,\ \beta_H \leq g_{\varepsilon}(\alpha;P) \mid \mathcal{D}_j)\,. \tag{2.13}$$

The rule is defined when the maximum in (2.13) is attained. Replace doctor j if $\widehat{q}^{\varepsilon}_{\max,j} \geq \tau$, using a threshold that achieves the maximizing point of the worst-case ROC curve, and retain the doctor if $\widehat{q}^{\varepsilon}_{\max,j} < \tau$.

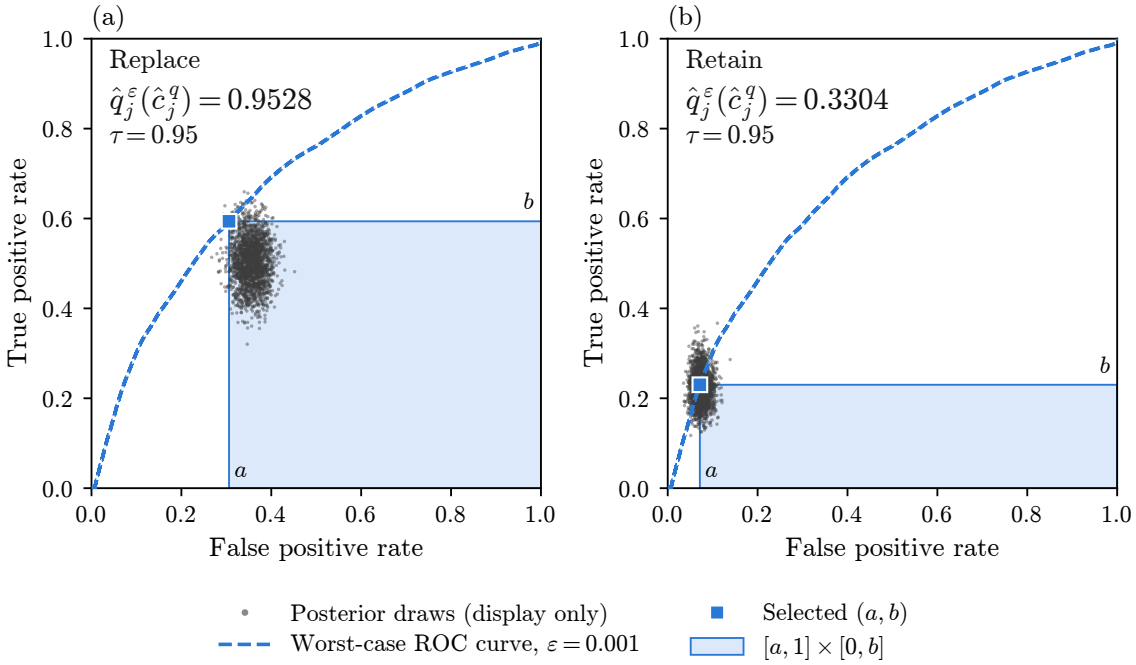


Figure 2: The posterior test at one threshold.

Notes: Two doctors in the first replication of Section 5, with the average budget $\varepsilon = 0.001$: (a) a doctor who is replaced and (b) a doctor who is retained. The dashed curve is the estimated worst-case ROC curve. The square marks the estimated worst-case FPR and TPR, a and b, of the selected threshold $\widehat{c}_j^{q}$ in (2.15), which has the highest posterior probability of dominance, and this threshold dominates every pair of error rates in the shaded rectangle $[a,1]\times[0,b]$. The posterior probability of the rectangle, $\widehat{q}_j^{\varepsilon}(\widehat{c}_j^{q})$, is computed exactly by (2.18); it is 0.9528 for the first doctor and 0.3304 for the second, so with $\tau = 0.95$ the first doctor is replaced and the second is retained. The dots are 2,000 draws from each doctor's posterior and are shown only for illustration.

The maximum in (2.13) exists, for instance, when the thresholds are restricted to a finite set of candidates. By Lemma 2, for each α the value $g_\varepsilon(\alpha; P)$ is attained at the smallest threshold that is feasible at level α, and

$$\sup_{\alpha\in[0,1]} \Pi_j\{\alpha_H \geq \alpha,\ \beta_H \leq g_\varepsilon(\alpha; P) \mid \mathcal{D}_j\} = \sup_{c\in[0,1]} q_j^\varepsilon(c; P).$$

Whenever the maximum exists, the threshold of Definition 3 can therefore be found by maximizing (2.12) over c. The same holds when the worst-case rates are computed exactly from $\widehat{\mathbb{P}}_n$, and for a finite grid of thresholds that contains $c = 1$, provided the worst-case ROC curve is computed over the same grid.

With $P = P_0$, (2.12) is the posterior probability that c dominates the doctor under every plausible deployment distribution. We call the worst-case rates computed with $P = \widehat{\mathbb{P}}_n$ the estimated worst-case rates, and denote the threshold selected with them by $\widehat{c}_j$. In (2.12), only the doctor's error rates are random; we treat the estimated worst-case rates as known. Section 3 gives the order of the estimation error in these rates and shows how it enters the replacement and welfare bounds, and Appendix A.4 shows how to use simultaneous confidence bounds on the worst-case rates when they are available.

The replacement decision can be read as the solution of a simple decision problem. The planner prefers to replace the doctor, which frees the doctor for other work, but only if the posterior probability that the algorithm dominates the doctor under every distribution in the set is at least τ. With a replacement indicator $r_j \in \{0, 1\}$, the replacement decision solves

$$\max_{r_j\in\{0,1\},\, c\in[0,1]} r_j \quad \text{subject to} \quad r_j\{1 - q_j^\varepsilon(c; P)\} \leq 1 - \tau. \tag{2.14}$$

Retaining the doctor ($r_j = 0$) always satisfies the constraint, and replacing the doctor ($r_j = 1$) satisfies it exactly when the threshold c passes the posterior test $q_j^\varepsilon(c; P) \geq \tau$. Because the objective depends only on r_j, replacing the doctor with any threshold that passes the test solves (2.14). Under Definition 3, the planner chooses the one with the highest posterior probability of dominance.

For a fixed threshold, the posterior probability (2.12) is the probability of dominance under all distributions in the set jointly (Section 2.2), and it is no greater than the infimum over the set of the posterior probabilities of dominance under the individual distributions. Maximizing the latter over thresholds would give a maxmin choice based on posterior dominance probabilities, related to maxmin evaluation under ambiguity (Gilboa and Schmeidler, 1989). The two probabilities coincide when one distribution attains both worst-case rates, as under the pointwise budget of Section 4 (Proposition 3). Under an average budget, the joint requirement can be stricter.

A planner who puts weights $\phi \geq 0$ on the TPR and $\eta \geq 0$ on the FPR obtains welfare $\phi\,\mathrm{TPR}_Q(c) - \eta\,\mathrm{FPR}_Q(c)$ from threshold c under a distribution Q. Such a planner can

instead choose, among the thresholds that pass the test, the one with the highest lower bound on this welfare over the set. To state the two threshold choices with the estimated worst-case rates, let $\mathcal{C}_n$ be the set of candidate thresholds and write $\widehat{q}^{\varepsilon}_j(c) = q^{\varepsilon}_j(c; \widehat{\mathbb{P}}_n)$. Define

$$\begin{aligned} \mathcal{C}^{\tau}_j &= \{c \in \mathcal{C}_n : \widehat{q}^{\varepsilon}_j(c) \geq \tau\}, \\ \widehat{V}_{\varepsilon}(c) &= \phi \underline{\beta}_{\varepsilon}(c; \widehat{\mathbb{P}}_n) - \eta \overline{\alpha}_{\varepsilon}(c; \widehat{\mathbb{P}}_n), \\ \widehat{c}^{q}_j \in \arg\max_{c \in \mathcal{C}_n} \widehat{q}^{\varepsilon}_j(c), &\qquad \widehat{c}^{W}_j \in \arg\max_{c \in \mathcal{C}^{\tau}_j} \widehat{V}_{\varepsilon}(c). \end{aligned} \tag{2.15}$$

The first choice, $\widehat{c}^{q}_j$, is the threshold $\widehat{c}_j$ selected by Definition 3, and we call the second, $\widehat{c}^{W}_j$, the welfare choice. Here $\widehat{V}_{\varepsilon}(c)$ is the welfare lower bound: because it combines the lowest TPR and the highest FPR over $\mathcal{B}_{\varepsilon}(\widehat{\mathbb{P}}_n)$, the welfare of c under every distribution in $\mathcal{B}_{\varepsilon}(\widehat{\mathbb{P}}_n)$ is at least $\widehat{V}_{\varepsilon}(c)$. The two rates need not come from the same distribution, so the lowest welfare over $\mathcal{B}_{\varepsilon}(\widehat{\mathbb{P}}_n)$ can be higher than $\widehat{V}_{\varepsilon}(c)$. If $\mathcal{C}^{\tau}_j$ is empty, the doctor is retained. Otherwise both choices replace the doctor, so they replace the same doctors and differ only in the threshold they use.[2] For each replaced doctor,

$$\widehat{q}^{\varepsilon}_j(\widehat{c}^{q}_j) \geq \widehat{q}^{\varepsilon}_j(\widehat{c}^{W}_j) \geq \tau, \qquad \widehat{V}_{\varepsilon}(\widehat{c}^{W}_j) \geq \widehat{V}_{\varepsilon}(\widehat{c}^{q}_j). \tag{2.16}$$

We use $\widehat{c}^{q}_j$ as the main choice because it does not require the costs of the two errors. Feng et al. (2025) also discuss choosing the point on the ROC curve by minimizing a posterior expected loss, but leave general costs aside because specifying them requires domain knowledge. The welfare choice $\widehat{c}^{W}_j$ uses such costs, but only to choose among thresholds that already pass the posterior test.

We call the rule with $\varepsilon = 0$ the nominal rule and the rule with $\varepsilon > 0$ the robust rule. Since $g_{\varepsilon} \leq g$, the robust rule is more conservative than the nominal rule at the same cutoff τ. Figure B.1 in Appendix B.2 shows how quickly the worst-case ROC curve falls as the budget increases.

Proposition 2 (Conservativeness of robust replacement)**.** *For any $\varepsilon > 0$, the doctors replaced by the robust rule are a subset of those replaced by the nominal rule:*

$$\left\{j : \widehat{q}^{\varepsilon}_{\max,j} \geq \tau\right\} \subseteq \left\{j : \widehat{q}^{0}_{\max,j} \geq \tau\right\}.$$

The planner must choose the budget. It is measured in the units of X: when X is standardized, a budget of 0.1 lets X move by 0.1 on average, so a few cases may move much farther. Even a small budget can move the worst-case rates by an amount of order $\sqrt{\varepsilon}$ (Remark 1). The planner may want the set to contain the distribution of data from another hospital. Whether it does depends on how far this distribution is from

[2] Maximizers are chosen measurably; on a finite grid, ties are broken in favor of the smallest threshold.

the training distribution under $\mathcal{W}_\kappa$. Because $\mathcal{W}_\kappa$ keeps every label fixed, two difficulties arise. First, two samples with different prevalences are infinitely far apart, since moving X cannot change the prevalence. Second, the marginal W_1 distance, which compares the distributions of X with the two classes pooled, cannot take the place of $\mathcal{W}_\kappa$. It lets mass move from one class to the other and is therefore never larger than $\mathcal{W}_\kappa$. A value below the budget does not show that the other distribution lies in the set. For example, when $p = 1/2$ and the healthy and sick cases swap their distributions of X, the marginal W_1 distance is zero, but the algorithm's TPR and FPR are swapped as well. For these reasons we report results over a grid of budgets, which shows how the decision changes as the set grows.

The argument of Proposition 2 also shows that a doctor replaced at some budget is replaced at every smaller budget. Each doctor replaced by the nominal rule therefore has a critical budget: the doctor is replaced below it and retained above it. The planner can compare this budget with the shifts considered plausible.

2.4 Computation

We compute the rule from the validation cases and the doctors' histories. Write $\underline{\widehat{\beta}}_\varepsilon(c) = \underline{\beta}_\varepsilon(c; \widehat{\mathbb{P}}_n)$ and $\overline{\widehat{\alpha}}_\varepsilon(c) = \overline{\alpha}_\varepsilon(c; \widehat{\mathbb{P}}_n)$ for the estimated worst-case rates. The estimated worst-case ROC curve $g_\varepsilon(\cdot; \widehat{\mathbb{P}}_n)$ is obtained from them as in (2.9).

For doctor j, let $(\mathrm{TP}_j, \mathrm{FP}_j, \mathrm{FN}_j, \mathrm{TN}_j)$ be the numbers of true positives, false positives, false negatives and true negatives in the doctor's history, the four cells of the doctor's confusion matrix. Under the symmetric Dirichlet prior with parameter $a_0 > 0$ on each cell, the doctor's FPR and TPR are independent under the posterior (Lemma 6), and

$$\begin{aligned} \alpha_H \mid \mathcal{D}_j &\sim \mathrm{Beta}(a_0 + \mathrm{FP}_j, a_0 + \mathrm{TN}_j), \\ \beta_H \mid \mathcal{D}_j &\sim \mathrm{Beta}(a_0 + \mathrm{TP}_j, a_0 + \mathrm{FN}_j). \end{aligned} \tag{2.17}$$

Let $F_{\alpha,j}$ and $F_{\beta,j}$ denote the posterior cumulative distribution functions (CDFs) of α_H and β_H, and let $S_{\alpha,j} = 1 - F_{\alpha,j}$. With the estimated worst-case rates, the posterior probability (2.12) is therefore

$$\widehat{q}_j^{\,\varepsilon}(c) = S_{\alpha,j}\left(\overline{\widehat{\alpha}}_\varepsilon(c)\right) F_{\beta,j}\left(\underline{\widehat{\beta}}_\varepsilon(c)\right). \tag{2.18}$$

The Beta distributions are continuous, so (2.18) is the same whether the dominance inequalities are weak or strict.

We use a finite grid of candidate thresholds. The grid contains $c = 1$, which classifies every case as negative; its worst-case FPR is zero, so some threshold on the grid is feasible at every level α. We maximize (2.18) over the grid and replace the doctor if the maximum is at least τ, using the maximizing threshold. Because (2.18) is in closed form, no posterior simulation is needed, and the remaining errors come from estimating the worst-case rates,

from computing them numerically and from searching over a grid.[3]

Exact computation for the logistic score. For $\varepsilon > 0$, the estimated worst-case rates come from Proposition 1 with $P = \widehat{\mathbb{P}}_n$ and require the distances $d_c^{\pm}(X_i)$. These depend on the score. For the logistic score, which we use in Sections 5 and 6, they have a closed form; for the random forest in Section 6, we use the pointwise budget of Section 4.1 instead. Let $m(x) = \{1 + \exp[-(w_0 + x^\top w)]\}^{-1}$ on $\mathcal{X} = \mathbb{R}^d$, with $w \neq 0$, and define $z(x) = w_0 + x^\top w$ and $k_c = \log\{c/(1-c)\}$ for $0 < c < 1$. The distances in Proposition 1 are then

$$d_c^-(x) = \frac{(z(x) - k_c)_+}{\|w\|_2}, \qquad d_c^+(x) = \frac{(k_c - z(x))_+}{\|w\|_2}. \tag{2.19}$$

These are the Euclidean distances from x to the half-space $\{x' : z(x') \leq k_c\}$, where the algorithm classifies a case as negative, and to the half-space $\{x' : z(x') > k_c\}$, where it classifies a case as positive. The worst- and best-case misclassification bounds of Shafieezadeh-Abadeh et al. (2015, 2019) for a fixed linear classifier use the same distances; here we apply them within each class. Ho-Nguyen and Wright (2023) use a label-preserving cost of the form (2.6) and show that the worst-case misclassification probability of a classifier is determined by the conditional value-at-risk of the distance to misclassification. If X is restricted to a subset of $\mathbb{R}^d$, the distances must be computed within that subset.

For $\varepsilon > 0$ and $P = \widehat{\mathbb{P}}_n$, the dual objectives in (2.10) and (2.11) are sums over the n validation cases, so each evaluation takes $O(n)$ time. Because they are piecewise linear in λ, it suffices to evaluate them at 0, at $1/\varepsilon$, and at those reciprocals of the positive distances that lie in $[0, 1/\varepsilon]$. Equivalently, each estimated worst-case rate can be computed as in Example 1 (Lemma 3). For the TPR, sort the sick cases classified as positive by their distance to the threshold and move them onto it, closest first. Moving case i costs d_i/n, where d_i is its distance and n counts the validation cases of both classes. The cases are moved until the budget is used up, with only part of the last case moved if needed, or until no sick case is left above the threshold. For the FPR, the healthy cases classified as negative are moved across the threshold in the same way. A healthy case moved onto the threshold is still classified as negative, so it must move slightly farther, and the worst-case FPR is a supremum that these moves approach as the extra distance shrinks but need not attain.

Remark 1 (The scale of the average budget)**.** *For the logistic score in* (2.19)*, let* $S = z(X)/\|w\|_2$ *and* $a_c = k_c/\|w\|_2$*, so that a case is classified as positive when* $S > a_c$*, and its distance to the boundary is* $|S - a_c|$*. Write* H_y *for the CDF of* S *given* $Y = y$*, and let* $\alpha_0(c) = \overline{\alpha}_0(c; P)$ *and* $\beta_0(c) = \underline{\beta}_0(c; P)$ *denote the nominal rates. At a fixed interior*

[3]Probabilities close to zero, and the complements of probabilities close to one, are compared on the logarithmic scale.

threshold where both CDFs have continuous positive densities, as $\varepsilon \downarrow 0$,

$$\overline{\alpha}_\varepsilon(c) - \alpha_0(c) = \sqrt{\frac{2\varepsilon H_0'(a_c)}{1-p}} + o(\sqrt{\varepsilon}), \tag{2.20}$$

$$\beta_0(c) - \underline{\beta}_\varepsilon(c) = \sqrt{\frac{2\varepsilon H_1'(a_c)}{p}} + o(\sqrt{\varepsilon}). \tag{2.21}$$

As in Section 2.2, the mass closest to the boundary is moved first. Within a class, a strip of width s *next to the boundary has mass* $H_y'(a_c)s + o(s)$, *and moving it across costs* $H_y'(a_c)s^2/2 + o(s^2)$. *When the whole budget is spent on one class, the cases of that class can move by* ε/p *on average if they are sick and by* $\varepsilon/(1-p)$ *if they are healthy (Section 2.2). Setting the cost of the strip equal to this amount gives* s, *and the mass of the strip gives (2.20) and (2.21). Example 2 in Section 3.3 gives the exact calculation for uniform densities: with* $p = 1/2$, *both worst-case rates move by* $2\sqrt{\varepsilon}$.

For small ε, $\sqrt{\varepsilon}$ *is much larger than* ε. *Spending the budget on the cases closest to the threshold therefore moves far more probability across it than moving every case a distance* ε *toward the threshold, which uses the same budget. The expansions require densities, which* $\widehat{\mathbb{P}}_n$ *does not have. Under* $\widehat{\mathbb{P}}_n$, *each increase in the budget is spent on one case at a time, so the estimated worst-case rates are piecewise linear in* $\varepsilon > 0$.

3 Statistical theory

We first bound the estimation error in the worst-case rates and in the worst-case ROC curve. We then turn to the implemented rule, which uses the estimated worst-case rates, and to its welfare.

3.1 Estimating the worst-case ROC curve

We estimate the worst-case rates $\overline{\alpha}_\varepsilon(c; P_0)$ and $\underline{\beta}_\varepsilon(c; P_0)$, which are taken over $\mathcal{B}_\varepsilon(P_0)$, and the worst-case ROC curve $g_\varepsilon(\cdot; P_0)$. Recall that $p_0 = P_0(Y = 1)$, and let $\widehat{p} := n^{-1}\sum_{i=1}^n \mathbb{I}(Y_i = 1)$ denote the prevalence in the validation sample. For $\varepsilon > 0$, on samples with $0 < \widehat{p} < 1$ and at thresholds with two nonempty decision regions, Proposition 1 with $P = \widehat{\mathbb{P}}_n$ gives the estimated worst-case TPR

$$\widehat{\underline{\beta}}_\varepsilon(c) := \sup_{\lambda \geq 0} \left\{ \frac{1}{\widehat{p}} \frac{1}{n} \sum_{i=1}^n Y_i \min\{1, \lambda d_c^-(X_i)\} - \frac{\lambda}{\widehat{p}}\varepsilon \right\}, \tag{3.1}$$

and the estimated worst-case FPR

$$\widehat{\overline{\alpha}}_\varepsilon(c) := \inf_{\lambda \geq 0} \left\{ \frac{\lambda}{1-\widehat{p}}\varepsilon + \frac{1}{1-\widehat{p}} \frac{1}{n} \sum_{i=1}^n (1 - Y_i)\big(1 - \lambda d_c^+(X_i)\big)_+ \right\}. \tag{3.2}$$

At $\varepsilon = 0$ and for constant rules, we compute the estimated worst-case rates as in Section 2.2. We write $\widehat{g}_\varepsilon$ for the estimated worst-case ROC curve $g_\varepsilon(\cdot; \widehat{\mathbb{P}}_n)$ of Section 2.4:

$$\widehat{g}_\varepsilon(\alpha) := \sup \left\{ \widehat{\underline{\beta}}_\varepsilon(c) : \widehat{\overline{\alpha}}_\varepsilon(c) \leq \alpha \right\}.$$

With $P = P_0$, (2.10) and (2.11) contain expectations under P_0 of the following functions of a case (x, y), which we call the dual functions:

$$\psi^+_{\lambda,c}(x, y) := \frac{y}{p_0} \min\{1, \lambda d^-_c(x)\},$$

$$\psi^-_{\lambda,c}(x, y) := \frac{1 - y}{1 - p_0}(1 - \lambda d^+_c(x))_+.$$

The estimates (3.1) and (3.2) use averages over the validation cases in place of these expectations, and $\widehat{p}$ in place of p_0. By Proposition 1, the dual optimizations under P_0 and under $\widehat{\mathbb{P}}_n$ can both be restricted to $\Lambda := [0, 1/\varepsilon]$. We obtain uniform convergence under two conditions. First, the dual functions, indexed by λ and c, must form classes that are not too rich, so that their averages over the validation cases are close to their expectations uniformly in λ and c. Second, the worst-case FPR $\overline{\alpha}_\varepsilon(\cdot; P_0)$ must decrease in c at a rate bounded away from zero, so that a small error in the worst-case FPR moves the threshold that reaches a given FPR level only a little. Assumption 1 states the two conditions.

Assumption 1 (Complexity and regularity)**.** *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. The following conditions hold:*

(i) the classes $\mathcal{F}^\pm := \{\psi^\pm_{\lambda,c} : (\lambda, c) \in \Lambda \times \mathcal{C}\}$ are pointwise measurable and have bounded envelopes. For each class there are fixed finite constants K, v and a constant envelope M such that

$$\sup_H N(uM, \mathcal{F}^\pm, L_2(H)) \leq (K/u)^v, \qquad 0 < u < 1,$$

where H ranges over finitely supported probability distributions;

(ii) on $\mathcal{C}$, the worst-case rates under P_0 are continuously differentiable, with

$$0 < \underline{L}_\alpha \leq -\partial_c \overline{\alpha}_\varepsilon(c; P_0) \leq \overline{L}_\alpha < \infty, \qquad |\partial_c \underline{\beta}_\varepsilon(c; P_0)| \leq L_\beta < \infty.$$

For the FPR interval $[\alpha_l, \alpha_u] \subset (0, 1)$ under consideration, there is $b_C > 0$ such that

$$\overline{\alpha}_\varepsilon(c_l; P_0) > \alpha_u + b_C, \qquad \overline{\alpha}_\varepsilon(c_u; P_0) < \alpha_l - b_C.$$

The covering bound in part (i) is the usual polynomial bound for classes of Vapnik–Chervonenkis (VC) type.[4]

Remark 2 (A verifiable complexity condition for logistic scores)**.** *For the logistic score on $\mathbb{R}^d$ in (2.19), the exact dual functions are clipped affine functions of the fixed index $z(x)$. By Lemma 3, the classes $\mathcal{F}^{\pm}$ are then VC-subgraph and pointwise measurable, which gives part (i). Part (ii) also requires the stated derivative bounds (Remark 3 in the Appendix gives primitive conditions for them) and an FPR interval inside the range of the strictly decreasing worst-case FPR.*

For $\alpha \in [\alpha_l, \alpha_u]$, let c_α be the smallest threshold in $\mathcal{C}$ that is feasible at level α, and collect these thresholds in $\mathcal{C}_\alpha$:

$$c_\alpha = \min\{c \in \mathcal{C} : \overline{\alpha}_\varepsilon(c; P_0) \le \alpha\}, \qquad \mathcal{C}_\alpha = \{c_\alpha : \alpha \in [\alpha_l, \alpha_u]\}.$$

By the endpoint inequalities in Assumption 1(ii), the FPR interval $[\alpha_l, \alpha_u]$ lies strictly inside the range of $\overline{\alpha}_\varepsilon(\cdot; P_0)$ on $\mathcal{C}$. For each α in this interval, c_α is therefore also the smallest threshold in $[0, 1]$ that is feasible at level α, so restricting attention to $\mathcal{C}$ does not change the worst-case ROC curve on $[\alpha_l, \alpha_u]$. Once the estimation error is small, the same holds for the estimated curve, even though the estimated worst-case FPR may jump.

Theorem 1 (Uniform convergence of the worst-case ROC curve)**.** *Under Assumption 1, for fixed $\varepsilon > 0$:*

(i)

$$\sup_{c \in \mathcal{C}} \left| \widehat{\underline{\beta}}_\varepsilon(c) - \underline{\beta}_\varepsilon(c; P_0) \right| + \sup_{c \in \mathcal{C}} \left| \widehat{\overline{\alpha}}_\varepsilon(c) - \overline{\alpha}_\varepsilon(c; P_0) \right| = O_p\big(n^{-1/2}\big)\,;$$

(ii) on the FPR interval $[\alpha_l, \alpha_u]$ of Assumption 1(ii),

$$\sup_{\alpha \in [\alpha_l, \alpha_u]} |\widehat{g}_\varepsilon(\alpha) - g_\varepsilon(\alpha; P_0)| = O_p\big(n^{-1/2}\big)\,. \tag{3.3}$$

The replacement and welfare results below rely on this uniform accuracy. Under additional continuity conditions, we strengthen Theorem 1 to a functional limit theorem in Appendix A.2 and verify these conditions for the logistic score.

3.2 Replacement decisions and selected thresholds

Whether a doctor is replaced correctly depends on the estimated worst-case ROC curve and on the posterior for the doctor's error rates. We first show that, with probability

[4] A class is pointwise measurable if a countable subclass approximates every member pointwise. For a probability distribution H, $N(t, \mathcal{F}, L_2(H))$ is the minimum number of $L_2(H)$ balls of radius t needed to cover $\mathcal{F}$.

approaching one, the implemented rule makes the correct replacement decision for every doctor. We then bound the error rates of the selected threshold under every distribution in the set.

As in Section 2.3, $\mathcal{C}_n \subseteq [0,1]$ is the set of candidate thresholds. It contains $c = 1$ and is either a finite grid or the full interval $[0,1]$. In the rest of this section, $\widehat{g}_\varepsilon$ denotes the estimated worst-case ROC curve computed over $\mathcal{C}_n$, and

$$\widehat{c}_j \in \arg\max_{c \in \mathcal{C}_n} \widehat{q}_j^{\varepsilon}(c), \qquad \widehat{q}_{\max,j}^{\varepsilon} = \widehat{q}_j^{\varepsilon}(\widehat{c}_j). \tag{3.4}$$

The maximizer in (3.4) is chosen measurably, which on a finite grid any fixed tie-breaking rule achieves. When $\mathcal{C}_n$ is the full interval, we also require the maximum to be attained, as in Definition 3; with the Beta posteriors in (2.17), this holds if both estimated worst-case rates are continuous on the interval. By Lemma 2, $\widehat{q}_{\max,j}^{\varepsilon}$ is also the maximum in (2.13) computed with $\widehat{g}_\varepsilon$. On a grid, $\widehat{g}_\varepsilon$ is computed over the grid points, and its error relative to $g_\varepsilon(\cdot; P_0)$, which is computed over all thresholds, is part of the error terms below.

Assumption 2 (Separation and regularity)**.**

(i) The number of doctors J is fixed. There are $\delta_0 > 0$ and $d_H > 0$ such that $\alpha_{H,j} \in [\alpha_l + d_H, \alpha_u - d_H]$ for every j, and either

$$\beta_{H,j} \le g_\varepsilon(\alpha_{H,j}; P_0) - \delta_0 \quad \text{or} \quad \beta_{H,j} \ge g_\varepsilon(\alpha_{H,j}; P_0) + \delta_0.$$

(ii) On $[\alpha_l, \alpha_u] \subset (0,1)$, the worst-case ROC curve $\alpha \mapsto g_\varepsilon(\alpha; P_0)$ is Lipschitz with constant $L_g \ge 0$.

Part (i) requires each doctor's FPR to lie inside the FPR interval $[\alpha_l, \alpha_u]$, away from its ends, and each doctor's TPR to differ by at least δ_0 from the worst-case ROC curve at the doctor's FPR. No doctor is therefore close to the boundary between being dominated and not being dominated. Because J is fixed, such δ_0 and d_H exist whenever no doctor lies exactly on the worst-case ROC curve and every doctor's FPR lies in the open interval (α_l, α_u). Under Assumption 1, part (ii) holds with $L_g = L_\beta / \underline{L}_\alpha$.

For $t > 0$, let $T_j(t)$ be the posterior probability that the doctor's error rates are more than t away from their true value $\theta_{H,j}$ in the maximum norm:

$$T_j(t) := \Pi_j(\|\theta_H - \theta_{H,j}\|_\infty > t \mid \mathcal{D}_j)\,. \tag{3.5}$$

We call $T_j(t)$ the posterior tail. It is random because the posterior depends on the history $\mathcal{D}_j$.

Assumption 3 (Mean posterior concentration)**.** *For each doctor with a history of $n_j \geq 1$ cases, there are constants $C_j \geq 1$ and $\zeta_j > 0$, independent of n_j and t, such that*

$$\mathbb{E}_{P_0}[T_j(t)] \leq C_j \exp(-\zeta_j n_j t^2), \qquad 0 < t \leq 1. \tag{3.6}$$

The expectation is over histories $\mathcal{D}_j$ drawn repeatedly from P_0, with the doctor's decision process held fixed.

We write $C_0 = \max_j C_j$, $\zeta_0 = \min_j \zeta_j$ and $n_{\min} = \min_j n_j$. Because (3.6) restricts only an expectation, an unusual history may still produce a posterior far from the truth. By Lemma 6, (3.6) holds when the history consists of independent cases, the prevalence is bounded away from zero and one, and the prior is the symmetric Dirichlet prior of Section 2.4. The results below combine the doctors through union bounds, so they do not require independence across doctors or between the histories and the validation sample.

For a logistic score with the Dirichlet posterior of Section 2.4, Assumption 1, Assumption 2(ii) and Assumption 3 therefore follow from primitive conditions (Remarks 2 and 3 and Lemma 6).

Let

$$r_n := \sup_{\alpha \in [\alpha_l, \alpha_u]} |\widehat{g}_\varepsilon(\alpha) - g_\varepsilon(\alpha; P_0)|$$

be the error of the estimated worst-case ROC curve on the FPR interval. A replacement decision is wrong if it differs from the decision we would make if we knew the doctor's true error rates and the worst-case ROC curve: replace the doctor if $\beta_{H,j} < g_\varepsilon(\alpha_{H,j}; P_0)$, and retain the doctor if $\beta_{H,j} > g_\varepsilon(\alpha_{H,j}; P_0)$. By Assumption 2(i), every doctor lies clearly on one side of the curve, so this decision is well defined.

Theorem 2 (Consistency of the replacement decisions)**.** *Suppose that (2.4) holds, that thresholds are selected as in (3.4), and that Assumptions 2 and 3 hold. Let*

$$0 < t_0 \leq \min\left\{ \frac{d_H}{2}, \frac{\delta_0}{4(1 + L_g)} \right\}.$$

On $\{r_n \leq \delta_0/4\}$:

(i) if $\beta_{H,j} \leq g_\varepsilon(\alpha_{H,j}; P_0) - \delta_0$, then $\widehat{q}^{\varepsilon}_{\max,j} \geq 1 - T_j(t_0)$;

(ii) if $\beta_{H,j} \geq g_\varepsilon(\alpha_{H,j}; P_0) + \delta_0$, then $\widehat{q}^{\varepsilon}_{\max,j} \leq T_j(t_0)$.

Thus $\widehat{q}^{\varepsilon}_{\max,j} \to_p 1$ or 0, respectively, when $r_n \to_p 0$ and $n_{\min} \to \infty$. For $\bar{\tau} = \min\{\tau, 1 - \tau\}$, the probability that at least one of the J doctors is assigned the wrong replacement decision obeys

$$\mathbb{P}(\textit{at least one wrong decision}) \leq \mathbb{P}(r_n > \delta_0/4) + \frac{JC_0}{\bar{\tau}} \exp(-\zeta_0 n_{\min} t_0^2). \tag{3.7}$$

The first term in (3.7) comes from the error in the estimated worst-case ROC curve, and the second comes from the posteriors for the doctors' error rates. For a dominated doctor, the proof finds a single threshold that dominates every pair of error rates within t_0 of the doctor's true error rates. For a doctor who is not dominated, no candidate threshold dominates any such pair. Because r_n includes any error from the grid or from numerical approximation, consistency also requires the grid to be fine enough; a fixed coarse grid may not suffice.

Write

$$\begin{aligned} e_{\alpha,n} &:= \sup_{c\in\mathcal{C}_n} \left|\widehat{\overline{\alpha}}_\varepsilon(c) - \overline{\alpha}_\varepsilon(c;P_0)\right|, \\ e_{\beta,n} &:= \sup_{c\in\mathcal{C}_n} \left|\widehat{\underline{\beta}}_\varepsilon(c) - \underline{\beta}_\varepsilon(c;P_0)\right| \end{aligned} \tag{3.8}$$

for the largest errors in the estimated worst-case rates over the candidate thresholds and, under Assumption 3, for $\gamma \in (0,1)$, put

$$t_n(\gamma) := \min\left\{1, \sqrt{\frac{\log(2JC_0/(\gamma\tau))}{\zeta_0 n_{\min}}}\right\}. \tag{3.9}$$

Theorem 3 (Error rates of the selected threshold)**.** *Suppose (2.4) holds, and let $\widehat{c}_j$ be a selected candidate with $\widehat{q}_j^{\varepsilon}(\widehat{c}_j) \geq \tau$. For every $Q \in \mathcal{B}_\varepsilon(P_0)$,*

$$\begin{aligned} \mathrm{FPR}_Q(\widehat{c}_j) &\leq \widehat{\overline{\alpha}}_\varepsilon(\widehat{c}_j) + e_{\alpha,n}, \\ \mathrm{TPR}_Q(\widehat{c}_j) &\geq \widehat{\underline{\beta}}_\varepsilon(\widehat{c}_j) - e_{\beta,n}. \end{aligned} \tag{3.10}$$

Consequently, under the posterior Π_j, the event

$$\mathrm{FPR}_Q(\widehat{c}_j) \leq \alpha_H + e_{\alpha,n}, \qquad \mathrm{TPR}_Q(\widehat{c}_j) \geq \beta_H - e_{\beta,n} \quad \textit{for every } Q \in \mathcal{B}_\varepsilon(P_0)$$

has probability at least τ. For any $t > 0$ with $T_j(t) < \tau$, the doctor's true error rates satisfy

$$\mathrm{FPR}_Q(\widehat{c}_j) \leq \alpha_{H,j} + e_{\alpha,n} + t, \qquad \mathrm{TPR}_Q(\widehat{c}_j) \geq \beta_{H,j} - e_{\beta,n} - t \tag{3.11}$$

for every $Q \in \mathcal{B}_\varepsilon(P_0)$. Under Assumption 3, for every $\gamma \in (0,1)$, with sampling probability at least $1-\gamma$, (3.11) holds with $t = t_n(\gamma)$ for every replaced doctor and every $Q \in \mathcal{B}_\varepsilon(P_0)$.

Theorem 3 bounds how much worse the selected threshold can be than the doctor it replaces, simultaneously for every distribution in the set. The first part holds under the posterior: with posterior probability at least τ, the selected threshold dominates the doctor under every distribution in the set, up to the estimation errors $e_{\alpha,n}$ and $e_{\beta,n}$. The second part holds under repeated sampling of the histories. It concerns the doctor's true error rates and therefore adds $t_n(\gamma)$, which shrinks as the histories grow.

3.3 Welfare comparisons

We compare the welfare of the implemented rule under a deployment distribution Q with that of two alternatives. The first is to retain all doctors, that is, not to use the algorithm. The welfare difference is the gain over retention. The second is the matched-FPR benchmark, a rule that knows Q and the doctors' error rates and so cannot be implemented. We consider the two in turn.

As in Section 2.3, we put weights $\phi \geq 0$ on the TPR and $\eta \geq 0$ on the FPR.[5] Under Q, the welfare of the algorithm at threshold c and the welfare of doctor j are

$$W_M(c; Q) := \phi \,\mathrm{TPR}_Q(c) - \eta \,\mathrm{FPR}_Q(c), \qquad W_{H,j} := \phi\beta_{H,j} - \eta\alpha_{H,j}.$$

By (2.4), the doctor's welfare is the same under every Q. If the doctors in a set $\mathcal{R}$ are replaced using a threshold profile $c = (c_j)_{j\in\mathcal{R}}$ and the other doctors are retained, the aggregate welfare is

$$W(\mathcal{R}, c; Q) := \frac{1}{J}\sum_{j\in\mathcal{R}} W_M(c_j; Q) + \frac{1}{J}\sum_{j\notin\mathcal{R}} W_{H,j}. \tag{3.12}$$

Retaining all doctors corresponds to $\mathcal{R} = \emptyset$.

Comparison with retaining all doctors. The implemented rule replaces the doctors in

$$\mathcal{R}_\varepsilon^{(n)} := \{j : \widehat{q}^{\,\varepsilon}_{\max,j} \geq \tau\}$$

and uses the threshold profile $\widehat{c} = (\widehat{c}_j)_{j\in\mathcal{R}_\varepsilon^{(n)}}$, where $\widehat{c}_j$ is selected by (3.4). Theorem 3 implies the following welfare guarantee.

Corollary 1 (Welfare relative to retaining all doctors)**.** *Suppose (2.4) and Assumption 3 hold, and doctor j is replaced exactly when $\mathcal{C}_j^\tau$ in (2.15) is nonempty, with a measurable threshold in that set. Let $\phi, \eta \geq 0$ and $\gamma \in (0,1)$. With sampling probability at least $1-\gamma$, simultaneously for every $Q \in \mathcal{B}_\varepsilon(P_0)$,*

$$W(\mathcal{R}_\varepsilon^{(n)}, \widehat{c}; Q) - W(\emptyset, \varnothing; Q) \geq -\frac{|\mathcal{R}_\varepsilon^{(n)}|}{J}\left[\phi e_{\beta,n} + \eta e_{\alpha,n} + (\phi+\eta)t_n(\gamma)\right], \tag{3.13}$$

where $e_{\alpha,n}$, $e_{\beta,n}$ and $t_n(\gamma)$ are defined in (3.8) and (3.9).

By (3.13), under every distribution in the set, replacement lowers welfare by at most the share of doctors replaced times the largest welfare loss per replaced doctor that (3.11) allows. If the worst-case rates and the doctors' error rates were known, this loss would

[5] If payoffs are specified per case, the prevalence, which is the same under every distribution in the set, is absorbed into ϕ and η.

be zero, because each replaced doctor would be dominated under every distribution in the set. Corollary 1 shows that the implemented rule, which uses estimated worst-case rates and a posterior, keeps this property up to terms that shrink with the errors in the estimated worst-case rates and with $t_n(\gamma)$. How much replacement gains depends on how much better the algorithm is than the replaced doctors under Q, and Section 5 reports these gains in simulations. For the nominal rule, $\varepsilon = 0$ and the set contains only P_0, so the guarantee covers only the training distribution.

Comparison with the matched-FPR benchmark. Corollary 1 does not show what the implemented rule gives up. To measure this, we compare it with the matched-FPR benchmark. Because the benchmark knows Q and the doctors' error rates, it faces no uncertainty and needs no worst-case rates. It replaces the doctors who lie below the algorithm's ROC curve under Q,

$$\mathcal{R}^*(Q) := \{j : \beta_{H,j} < g_Q(\alpha_{H,j})\},$$

and, for each $j \in \mathcal{R}^*(Q)$, uses a threshold $c_j^*(Q)$ at which the algorithm has the doctor's FPR under Q and the highest TPR at that FPR, $\mathrm{FPR}_Q(c_j^*(Q)) = \alpha_{H,j}$ and $\mathrm{TPR}_Q(c_j^*(Q)) = g_Q(\alpha_{H,j})$. We write $c^*(Q) := (c_j^*(Q))_{j \in \mathcal{R}^*(Q)}$. Such thresholds need not exist, for instance when FPR_Q jumps over $\alpha_{H,j}$. We say that the benchmark is defined at Q if they exist for every $j \in \mathcal{R}^*(Q)$, and we compare the implemented rule with the benchmark only at such Q.

We also use a population version of the robust rule. Like the benchmark, it cannot be implemented: it is the rule we would use if we knew the doctors' error rates and the worst-case rates under P_0, but not Q, and its replacement decisions are those that Theorem 2 calls correct. It replaces the doctors who lie below the worst-case ROC curve under P_0,

$$\mathcal{R}_\varepsilon(P_0) := \{j : \beta_{H,j} < g_\varepsilon(\alpha_{H,j}; P_0)\},$$

and, for each, uses the threshold whose worst-case FPR equals the doctor's FPR. If the doctor's error rates are known, the probability in (2.12) is one for every threshold that dominates the doctor, so it does not single out a threshold, and we match the doctor's FPR as the benchmark does. Definition 4 states both comparisons.

Definition 4 (Matched-FPR welfare comparison)**.** *For the implemented rule, define the regret*

$$\mathrm{Regret}(\varepsilon, Q) := W(\mathcal{R}^*(Q), c^*(Q); Q) - W(\mathcal{R}_\varepsilon^{(n)}, \widehat{c}; Q).$$

For a population comparison, suppose that each $j \in \mathcal{R}_\varepsilon(P_0)$ has a threshold $c_{\varepsilon,j}$ with

$$\overline{\alpha}_\varepsilon(c_{\varepsilon,j}; P_0) = \alpha_{H,j}, \qquad \underline{\beta}_\varepsilon(c_{\varepsilon,j}; P_0) = g_\varepsilon(\alpha_{H,j}; P_0).$$

Write $c_\varepsilon = (c_{\varepsilon,j})_{j \in \mathcal{R}_\varepsilon(P_0)}$ *and define the population regret*

$$\mathrm{Regret}_{\mathrm{pop}}(\varepsilon, Q) := W(\mathcal{R}^*(Q), c^*(Q); Q) - W(\mathcal{R}_\varepsilon(P_0), c_\varepsilon; Q). \tag{3.14}$$

The regret $\mathrm{Regret}(\varepsilon, Q)$ is the welfare that the implemented rule gives up relative to the benchmark. Unlike regret relative to the best rule, it can be negative, because the benchmark need not maximize welfare under Q. The population regret $\mathrm{Regret}_{\mathrm{pop}}(\varepsilon, Q)$ compares the benchmark with the population version of the robust rule. Neither of these two rules estimates anything, so the population regret measures only the price of guarding against shifts. The difference between the two regrets is the welfare difference between the population version and the implemented rule. Part of it comes from estimation and shrinks as the samples grow. The other part comes from the different choice of threshold and need not vanish; the threshold term defined below accounts for it.

To see where the regret comes from, we split the doctors replaced by the benchmark or by the implemented rule into those replaced by both (common replacements), those replaced only by the benchmark (missed replacements) and those replaced only by the implemented rule (extra replacements):

$$\mathcal{I}_n = \mathcal{R}^*(Q) \cap \mathcal{R}_\varepsilon^{(n)}, \quad \mathcal{M}_n = \mathcal{R}^*(Q) \setminus \mathcal{R}_\varepsilon^{(n)}, \quad \mathcal{A}_n = \mathcal{R}_\varepsilon^{(n)} \setminus \mathcal{R}^*(Q).$$

A common replacement contributes the welfare difference between the two thresholds. A missed replacement contributes the difference between the benchmark's threshold and the doctor, and an extra replacement contributes the difference between the doctor and the selected threshold. A doctor whom neither rule replaces contributes nothing. Hence

$$\begin{aligned} & W(\mathcal{R}^*, c^*; Q) - W(\mathcal{R}_\varepsilon^{(n)}, \widehat{c}; Q) \\ = & \frac{1}{J} \sum_{j \in \mathcal{I}_n} [W_M(c_j^*; Q) - W_M(\widehat{c}_j; Q)] \\ & + \frac{1}{J} \sum_{j \in \mathcal{M}_n} [W_M(c_j^*; Q) - W_{H,j}] \\ & + \frac{1}{J} \sum_{j \in \mathcal{A}_n} [W_{H,j} - W_M(\widehat{c}_j; Q)]. \end{aligned} \tag{3.15}$$

The first sum is present even when the two replacement sets coincide. Its terms can have either sign, because the benchmark need not maximize $\phi\,\mathrm{TPR}_Q - \eta\,\mathrm{FPR}_Q$ over all thresholds.

The worst-case ROC curve never lies above g_Q for Q in the set, so every doctor that the population version replaces is also replaced by the benchmark. Let

$$\Delta_\varepsilon(Q) := \sup_{\alpha \in [\alpha_l, \alpha_u]} \{g_Q(\alpha) - g_\varepsilon(\alpha; P_0)\} \geq 0$$

be the largest gap between the two curves on the FPR interval.

To bound the missed and extra replacements, we use the following condition, the counterpart of the margin conditions for classifiers (Mammen and Tsybakov, 1999; Audibert and Tsybakov, 2007) with doctors in place of cases.

Assumption 4 (Margin condition at a deployment distribution)**.** *At a deployment distribution $Q \in \mathcal{B}_\varepsilon(P_0)$, there are constants $C_M(Q), \gamma_M(Q) > 0$ such that*

$$\frac{1}{J}\sum_{j=1}^{J} \mathbb{I}(|g_Q(\alpha_{H,j}) - \beta_{H,j}| \le t) \le C_M(Q)\, t^{\gamma_M(Q)} \qquad \text{for every } t > 0. \tag{3.16}$$

The left side of (3.16) is the share of doctors whose TPR is within t of g_Q at the doctor's FPR, so the condition limits how many doctors lie close to the algorithm's ROC curve under Q. Unlike Assumption 2(i), which concerns the worst-case ROC curve under P_0, Assumption 4 concerns the algorithm's ROC curve under the deployment distribution at which welfare is evaluated, and the constants may depend on that distribution, on the doctors and on the budget. Because J is fixed, (3.16) requires each doctor's TPR to differ by at least $(JC_M(Q))^{-1/\gamma_M(Q)}$ from g_Q at the doctor's FPR:

$$\min_{j \le J} |g_Q(\alpha_{H,j}) - \beta_{H,j}| \ge (JC_M(Q))^{-1/\gamma_M(Q)} > 0.$$

A doctor whose TPR differs by $d > 0$ from g_Q at the doctor's FPR makes the left side of (3.16) at least $1/J$ at $t = d$, which gives this bound, and a doctor with $d = 0$ would violate (3.16) as $t \downarrow 0$. Hence g_Q passes through no doctor's error rates. The separation bound and its constants are specific to the deployment distribution at which welfare is evaluated.

The bound on the first sum in (3.15) uses a quantity that can be computed from the data. For a replaced doctor, the population version of the robust rule uses the point on the worst-case ROC curve at the doctor's FPR. Since this FPR is unknown, we take a median $a_{n,j}$ of its posterior, which is unique for the Beta posterior in (2.17), and use the point $(a_{n,j}, \widehat{g}_\varepsilon(a_{n,j}))$ on the estimated worst-case ROC curve as a reference. The implemented rule instead uses the selected threshold $\widehat{c}_j$. Define the threshold term

$$\widehat{K}_n := \frac{1}{J}\sum_{j \in \mathcal{R}_\varepsilon^{(n)}} \Big(\phi\{\widehat{g}_\varepsilon(a_{n,j}) - \underline{\widehat{\beta}}_\varepsilon(\widehat{c}_j)\} + \eta\{\widehat{\overline{\alpha}}_\varepsilon(\widehat{c}_j) - a_{n,j}\} \Big)_+ . \tag{3.17}$$

Since $\phi\,\underline{\widehat{\beta}}_\varepsilon(c) - \eta\,\widehat{\overline{\alpha}}_\varepsilon(c)$ is the welfare lower bound $\widehat{V}_\varepsilon(c)$ of (2.15), the threshold term

evaluated at a threshold profile $c = (c_j)$, written $\widehat{K}_n(c)$, is

$$\widehat{K}_n(c) = \frac{1}{J} \sum_{j \in \mathcal{R}_\varepsilon^{(n)}} \big(\phi \widehat{g}_\varepsilon(a_{n,j}) - \eta a_{n,j} - \widehat{V}_\varepsilon(c_j)\big)_+.$$

Each term is the shortfall, if any, of the welfare lower bound of the selected threshold below $\phi\, \widehat{g}_\varepsilon(a_{n,j}) - \eta\, a_{n,j}$, the same combination of the two rates at the reference point. The median only locates the reference point, and the rule itself does not use it.

Fix $\gamma \in (0,1)$, let $\bar{\tau} = \min\{\tau, 1-\tau\}$ as in Theorem 2, and let

$$t := \sqrt{\frac{\log\{2JC_0/(\gamma\bar{\tau})\}}{\zeta_0 n_{\min}}}. \tag{3.18}$$

With r_n as defined before Theorem 2, and with $e_{\alpha,n}$ and $e_{\beta,n}$ as in (3.8), define the nonnegative tolerances

$$\begin{aligned} h_n &:= r_n + (1 + L_g)t, \\ u_n &:= \phi e_{\beta,n} + \eta e_{\alpha,n} + (\phi + \eta)t, \\ v_n &:= \phi r_n + \phi e_{\beta,n} + \eta e_{\alpha,n} + (\phi L_g + \eta)t, \end{aligned}$$

which combine the estimation errors with t, and let $\widehat{\pi}_n := |\mathcal{R}_\varepsilon^{(n)}|/J$ be the share of doctors replaced.

Theorem 4 (Welfare relative to the matched-FPR benchmark)**.** *Suppose* (2.4) *holds and thresholds are selected by* (3.4)*. Let $\phi, \eta \geq 0$, and suppose that every doctor's FPR lies in $[\alpha_l + d_H, \alpha_u - d_H]$ for a fixed $d_H > 0$. Let $Q \in \mathcal{B}_\varepsilon(P_0)$ be a distribution at which the benchmark is defined and Assumption 4 holds, and write $C_M = C_M(Q)$ and $\gamma_M = \gamma_M(Q)$.*

(i) If the thresholds $c_{\varepsilon,j}$ in Definition 4 exist, then

$$\mathrm{Regret}_{\mathrm{pop}}(\varepsilon, Q) \leq \phi \frac{|\mathcal{R}_\varepsilon(P_0)|}{J} \Delta_\varepsilon(Q) + \phi C_M \Delta_\varepsilon(Q)^{1+\gamma_M}. \tag{3.19}$$

(ii) Suppose also that Assumptions 2(ii) and 3 hold and that $t \leq \min\{1, d_H\}$. There is an event, not depending on Q, with sampling probability at least $1 - \gamma$, on which

$$\begin{aligned} \mathrm{Regret}(\varepsilon, Q) \leq \widehat{K}_n &+ \widehat{\pi}_n\{\phi\Delta_\varepsilon(Q) + v_n\} \\ &+ \phi C_M\{\Delta_\varepsilon(Q) + h_n\}^{1+\gamma_M} + C_M u_n h_n^{\gamma_M} \end{aligned} \tag{3.20}$$

and the shares of extra and missed replacements obey

$$\frac{|\mathcal{A}_n|}{J} \leq C_M h_n^{\gamma_M}, \qquad \frac{|\mathcal{M}_n|}{J} \leq C_M\{\Delta_\varepsilon(Q) + h_n\}^{\gamma_M}. \tag{3.21}$$

On this event, (3.20) and (3.21) hold simultaneously for every $Q \in \mathcal{B}_\varepsilon(P_0)$ at which the benchmark is defined and Assumption 4 holds, with the constants associated with that Q. The statement also holds for the welfare choice $\widehat{c}^W$ in (2.15), with $\widehat{K}_n$ evaluated at $\widehat{c}^W$ and the same replacement set.

In part (ii), probabilities are over the histories and the validation sample, given the fitted score. The errors r_n, $e_{\alpha,n}$ and $e_{\beta,n}$ are random and unobserved, so (3.20) gives a numerical confidence bound on the regret only when combined with a confidence bound on these errors.

The bounds in Theorem 4 contain three kinds of terms: terms in $\Delta_\varepsilon(Q)$, the price of guarding against shifts; the tolerances h_n, u_n and v_n, which tend to zero as the estimation errors vanish and the histories grow; and the threshold term $\widehat{K}_n$, which comes from the choice of threshold. Part (i) concerns the population regret, which involves no estimation, so (3.19) contains only terms in $\Delta_\varepsilon(Q)$. They come from the common replacements, where the population version matches the doctor's FPR in the worst case rather than under Q, and from the missed replacements, the doctors between the two curves, whose number Assumption 4 limits. Part (ii) concerns the implemented rule. In (3.20), the first two terms come from the common replacements, the third from the missed replacements and the fourth from the extra replacements. Proposition 4 in the Appendix gives the stochastic order of the positive part of the regret after $\widehat{K}_n$ and $\phi\widehat{\pi}_n\Delta_\varepsilon(Q)$ are subtracted.

Unlike the tolerances, $\widehat{K}_n$ need not tend to zero, even when the estimation errors vanish and the histories grow. The rule ranks thresholds by the product of the two posterior probabilities in (2.18), rather than by welfare. Several strictly dominating thresholds may have posterior probabilities close to one as the history grows. This does not force the maximizer to approach a threshold whose worst-case FPR matches the doctor's FPR. Example 2 gives a symmetric setting in which the selected threshold remains separated from that threshold and $\widehat{K}_n$ stays positive.

Example 2 (A threshold term that does not vanish)**.** *Let $\mathcal{X} = \mathbb{R}$, $P_0(Y = 1) = 1/2$, and $X \mid Y = y \sim \mathrm{Unif}[0, 1]$ for both classes. Let $m(x) = (1 + e^{-x})^{-1}$, suppose there is one doctor, whose error rates $(\alpha_H, \beta_H) = (0.8, 0.2)$ satisfy (2.4), and let $(\phi, \eta) = (2, 1)$. Write $s = 2\sqrt{\varepsilon} \in (0, 0.1)$. At thresholds with $\log\{c/(1-c)\} \in [0, 1]$, the nominal FPR and TPR both equal $u = 1 - \log\{c/(1-c)\}$, so the ROC curve under P_0 is the diagonal, $g_{P_0}(\alpha) = \alpha$. For $s \le u \le 1 - s$, both worst-case rates move by exactly s, the leading term of (2.20) and (2.21):*

$$\overline{\alpha}_\varepsilon(c; P_0) = u + s, \qquad \underline{\beta}_\varepsilon(c; P_0) = u - s.$$

Hence $g_\varepsilon(\alpha; P_0) = \alpha - 2s$ for $\alpha \in [0.7, 0.9]$. At $Q = P_0$, the welfare of the algorithm at nominal rate u is $2u - u = u$. The matched-FPR benchmark uses $u = 0.8$, whereas the population version uses $u = 0.8 - s$, at which the worst-case FPR equals 0.8. Both replace

the doctor, yet

$$\text{Regret}_{\text{pop}}(\varepsilon, P_0) = s.$$

With the FPR interval $[\alpha_l, \alpha_u] = [0.7, 0.9]$, $\Delta_\varepsilon(P_0) = 2s$, *and Assumption 4 holds at every* Q *in the set, with* $\gamma_M(Q) = 1$ *and* $C_M(Q) = (0.6 - 2s)^{-1}$.

Now suppose that the worst-case rates under P_0 *are known, so that they contain no estimation error, and that the doctor's history consists of a growing number of independent cases, with the Dirichlet parameter* $a_0 > 0$ *held fixed. Every exact maximizer of the posterior probability (2.12) has nominal rate* $\widehat{u} \to 0.5$ *almost surely, and the doctor is eventually replaced for every fixed* $\tau < 1$. *The welfare of the selected threshold therefore tends to* 0.5, *whereas that of the benchmark is* 0.8, *and the implemented rule has*

$$\text{Regret}(\varepsilon, P_0) \longrightarrow 0.3, \qquad \widehat{K}_n \longrightarrow 0.3 - s.$$

The population regret is of the order of $\Delta_\varepsilon(P_0)$, *whereas the regret of the implemented rule contains a threshold term that does not vanish. In this symmetric example, the selected threshold balances the rates at which the posterior probabilities of* $\alpha_H < u + s$ *and of* $\beta_H > u - s$ *decay. These rates are equal when the two gaps,* $0.8 - (u + s)$ *and* $(u - s) - 0.2$, *are equal, which happens at* $u = 0.5$ *whatever the welfare weights. Appendix A.2 gives the calculations.*

If the welfare weights are specified, the welfare choice $\widehat{c}^W$ in (2.15) can be used instead. It never has a larger threshold term. The two choices replace the same doctors, the medians $a_{n,j}$ do not depend on the threshold, and by (2.16) the welfare choice has a welfare lower bound at least as high for each replaced doctor, so

$$\widehat{K}_n(\widehat{c}^W) \leq \widehat{K}_n(\widehat{c}^q). \tag{3.22}$$

This compares a term of the bound, not welfare itself; Section 5.5 compares the welfare of the two choices in simulations.

Theorems 2 and 3, Corollary 1 and Theorem 4(ii) also apply to $\widehat{c}^W$. The two choices make the same replacement decisions, so Theorem 2 applies, although its statement about $\widehat{q}^{\varepsilon}_{\max,j}$ concerns the largest posterior probability, not the posterior probability at $\widehat{c}^W_j$. The proofs of Theorem 3, Corollary 1 and Theorem 4(ii) use only two properties, which $\widehat{c}^W$ shares with $\widehat{c}^q$: the threshold used for each replaced doctor passes the posterior test, and a doctor is replaced whenever some candidate threshold passes it.

4 An extension to pointwise budgets

In Section 2.2, the budget ε limits only how far the cases move on average, so a few cases may move far if the others move little. In this section we instead limit how far each case

can move, a restriction related to the perturbations considered by Madry et al. (2017). Such a limit fits shifts that move each case by at most a known distance, for example after a change in measurement equipment. We call this limit the pointwise budget and denote it by ρ. We call ε the average budget. A pointwise budget needs only the lowest and highest scores within distance ρ of each case, so it also applies to tree-based scores such as random forests. Section 4.1 defines the worst-case ROC curve under a pointwise budget, computes it for logistic scores and approximates it for tree-based scores. Section 4.2 bounds the error of its estimate and shows that the replacement and welfare results of Section 3 carry over.

4.1 The worst-case ROC curve under a pointwise budget

Under a pointwise budget, each case may move by at most ρ, and its label does not change.

Definition 5 (Set under a pointwise budget)**.** *Let $\mathcal{X}$ be a closed subset of $\mathbb{R}^d$. For $\rho \geq 0$, write $\mathbb{B}_\rho(x) = \{z \in \mathcal{X} : \|z - x\|_2 \leq \rho\}$ and define*

$$\tilde{\mathcal{B}}_\rho(P) := \{(T_X, Id_Y)_\# P : T_X \text{ is measurable and } T_X(X,Y) \in \mathbb{B}_\rho(X) \quad P\text{-a.s.}\}.$$

The pushforward is the distribution of $(T_X(X,Y), Y)$.

The map T_X may depend on the label, so that a sick case and a healthy case at the same x can move in different directions.

The score m stays fixed: a shift changes the score of a case only by moving the case. A case at x can therefore end up with any score $m(z)$ with $z \in \mathbb{B}_\rho(x)$, and the worst-case rates depend on the lowest and highest of these scores.

Definition 6 (Local extrema of the score)**.** *For a fixed score $m : \mathcal{X} \to [0,1]$, define*

$$\underline{m}_\rho(x) := \inf_{z \in \mathbb{B}_\rho(x)} m(z), \tag{4.1}$$

$$\overline{m}_\rho(x) := \sup_{z \in \mathbb{B}_\rho(x)} m(z). \tag{4.2}$$

We use these extrema when they are measurable and are attained at measurable selections $z_\rho^-(x), z_\rho^+(x) \in \mathbb{B}_\rho(x)$.

These conditions hold when m is continuous, and when m is a fixed finite ensemble of axis-aligned trees on $\mathbb{R}^d$ whose splits assign boundary points to a specified side (Lemma 7).

At a threshold c, a sick case at x can be moved to a point that the algorithm classifies as negative exactly when $\underline{m}_\rho(x) \leq c$, and a healthy case at x can be moved to a point that the algorithm classifies as positive exactly when $\overline{m}_\rho(x) > c$. We call $\overline{m}_\rho(x)$ the worst-case score of a healthy case at x, and $\underline{m}_\rho(x)$ that of a sick case.

Definition 7 (Worst-case ROC curve under a pointwise budget)**.** *For* $0 < p = P(Y = 1) < 1$*, define the rates*

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

and the curve

$$\tilde{g}_\rho(\alpha; P) = \sup_{c \in [0,1]} \{\tilde{\beta}_\rho(c; P) : \tilde{\alpha}_\rho(c; P) \leq \alpha\}, \qquad \alpha \in [0, 1]. \tag{4.3}$$

Let $N_y = \sum_{i=1}^n \mathbb{I}(Y_i = y)$*. On samples with both classes, the curve under* $\widehat{\mathbb{P}}_n$ *is*

$$\tilde{g}_\rho(\alpha; \widehat{\mathbb{P}}_n) = \sup_{c \in [0,1]} \left\{ \frac{1}{N_1} \sum_{i:Y_i=1} \mathbb{I}(\underline{m}_\rho(X_i) > c) : \frac{1}{N_0} \sum_{i:Y_i=0} \mathbb{I}(\overline{m}_\rho(X_i) > c) \leq \alpha \right\}. \tag{4.4}$$

The threshold $c = 1$ classifies every case as negative and satisfies the constraint in (4.3) for every α, so the curve is well defined. As in (2.2), we do not choose at random between two thresholds.

Proposition 3 (Worst-case rates under a pointwise budget)**.** *Suppose the measurable attaining selections in Definition 6 exist,* P *has a finite first moment, and* $0 < P(Y = 1) < 1$*. Then, for every* c*,*

$$\inf_{Q \in \tilde{\mathcal{B}}_\rho(P)} \mathrm{TPR}_Q(c) = \tilde{\beta}_\rho(c; P), \tag{4.5}$$

$$\sup_{Q \in \tilde{\mathcal{B}}_\rho(P)} \mathrm{FPR}_Q(c) = \tilde{\alpha}_\rho(c; P). \tag{4.6}$$

Moreover, $\tilde{g}_0(\alpha; P) = g(\alpha; P)$*, and for* $0 \leq \rho \leq \varepsilon$*,*

$$g_\varepsilon(\alpha; P) \leq \tilde{g}_\rho(\alpha; P) \leq g(\alpha; P). \tag{4.7}$$

By (4.5) and (4.6), $\tilde{\beta}_\rho$ and $\tilde{\alpha}_\rho$ are the worst-case TPR and FPR over $\tilde{\mathcal{B}}_\rho(P)$, and $\tilde{g}_\rho$ is the worst-case ROC curve under the pointwise budget. Proposition 3 needs the extrema to be attained: without attainment, $\tilde{\beta}_\rho$ can be below the TPR under every distribution in $\tilde{\mathcal{B}}_\rho(P)$ (Appendix A.3). Moving each sick case to $z_\rho^-(x)$, where the score is lowest, and each healthy case to $z_\rho^+(x)$, where it is highest, gives a single distribution in $\tilde{\mathcal{B}}_\rho(P)$ that attains both worst-case rates at every threshold. One distribution can attain both worst-case rates because each case has its own limit ρ; under an average budget, the two worst cases need not come from the same distribution (Section 2.2).

Because one distribution attains both worst-case rates, for $\phi, \eta \geq 0$,

$$\inf_{Q \in \tilde{\mathcal{B}}_\rho(P)} \{\phi \mathrm{TPR}_Q(c) - \eta \mathrm{FPR}_Q(c)\} = \phi \tilde{\beta}_\rho(c; P) - \eta \tilde{\alpha}_\rho(c; P). \tag{4.8}$$

With $P = \widehat{\mathbb{P}}_n$, (4.8) shows that the welfare lower bound in (2.15), computed with $\tilde{\beta}_\rho$ and $\tilde{\alpha}_\rho$ in place of $\underline{\beta}_\varepsilon$ and $\overline{\alpha}_\varepsilon$, is the lowest welfare over $\tilde{\mathcal{B}}_\rho(\widehat{\mathbb{P}}_n)$. The welfare choice therefore maximizes this lowest welfare among the thresholds that pass the posterior test. Under an average budget, the welfare lower bound can be below the lowest welfare (Section 2.3).

The ordering (4.7) follows from $\{P\} \subseteq \tilde{\mathcal{B}}_\rho(P) \subseteq \mathcal{B}_\varepsilon(P)$ for $\rho \leq \varepsilon$: if every case moves by at most ρ, the cases move by at most ε on average. Under the attainment condition of Proposition 3, the same worst-case rates are obtained over the set of distributions reached by a label-preserving coupling that moves no mass farther than ρ, the ball of radius ρ in the Wasserstein distance of infinite order: allowing the mass of a case to split among several destinations cannot improve on the measurable maps that attain the local extrema of the score.

The logistic score. For the logistic score on $\mathbb{R}^d$ in (2.19), write $\sigma(t) = (1 + e^{-t})^{-1}$. By the Cauchy–Schwarz inequality and the monotonicity of σ,

$$\underline{m}_\rho(x) = \sigma\{z(x) - \rho\|w\|_2\}, \qquad \overline{m}_\rho(x) = \sigma\{z(x) + \rho\|w\|_2\}. \tag{4.9}$$

With a_c and H_y as in Remark 1,

$$\tilde{\alpha}_\rho(c) = 1 - H_0(a_c - \rho), \qquad \tilde{\beta}_\rho(c) = 1 - H_1(a_c + \rho). \tag{4.10}$$

By (4.10), at a fixed threshold, a pointwise budget moves the worst-case rates away from the nominal rates by an amount of order ρ, whereas an average budget moves them by an amount of order $\sqrt{\varepsilon}$ (Remark 1). When H_0 is continuous and strictly increasing at the relevant quantiles, set $a_\alpha = H_0^{-1}(1 - \alpha)$. The FPR constraint binds at $a_c = a_\alpha + \rho$, giving

$$\tilde{g}_\rho(\alpha) = 1 - H_1(a_\alpha + 2\rho), \qquad g(\alpha) - \tilde{g}_\rho(\alpha) = 2H_1'(a_\alpha)\rho + o(\rho). \tag{4.11}$$

The expansion holds when H_1' is continuous at a_α. The distribution that attains both worst-case rates comes from the map

$$T_\rho(x, y) = x + (1 - 2y)\rho\frac{w}{\|w\|_2}, \tag{4.12}$$

which moves each case a distance ρ parallel to w: sick cases in the direction that lowers the index $z(x)$, and healthy cases in the direction that raises it.

Tree-based scores. For tree-based scores on $\mathcal{X} = \mathbb{R}^d$, we approximate the local extrema by a search. We evaluate the score at the case itself and at K proposals, each equal to x plus a direction drawn uniformly from the unit sphere times an independent length

drawn from Unif$[0,\rho]$.[6] The smallest and largest scores found, $\underline{m}_{\rho,K}(x)$ and $\overline{m}_{\rho,K}(x)$, are used in place of $\underline{m}_\rho(x)$ and $\overline{m}_\rho(x)$ in the rates of Definition 7 and hence in the posterior probability (2.18), and the rule uses the threshold selected with them. Because the search stays in $\mathbb{B}_\rho(x)$ and includes x,

$$\underline{m}_\rho \le \underline{m}_{\rho,K} \le m \le \overline{m}_{\rho,K} \le \overline{m}_\rho.$$

For the same cases and thresholds, the curve computed from these scores lies between $\tilde{g}_\rho(\cdot;\widehat{\mathbb{P}}_n)$ and the ROC curve under $\widehat{\mathbb{P}}_n$. A finite search can therefore only understate the worst case, and this error need not vanish as the validation sample grows.

4.2 Estimation, replacement and welfare

Throughout this subsection, we fix a pointwise budget $\rho \ge 0$, and the local extrema are computed exactly and, like the score, held fixed. Let

$$Z_0(x) = \overline{m}_\rho(x), \qquad Z_1(x) = \underline{m}_\rho(x), \qquad F_y(c) = P_0(Z_y(X) \le c \mid Y = y),$$

so that $Z_y(x)$ is the worst-case score of a case at x with label y, and F_y is its CDF given $Y = y$. By Definition 7, $\tilde{\alpha}_\rho = 1 - F_0$ and $\tilde{\beta}_\rho = 1 - F_1$. The worst-case ROC curve under a pointwise budget is therefore the ROC curve obtained when each case is classified by its worst-case score, and it depends on P_0 only through F_0 and F_1.

We estimate F_y by the empirical CDF $\widehat{F}_y$ of the worst-case scores of the validation cases with label y. The estimated worst-case ROC curve is $\widehat{\tilde{g}}_\rho = \tilde{g}_\rho(\cdot;\widehat{\mathbb{P}}_n)$ in (4.4), which uses $\widehat{F}_0$ and $\widehat{F}_1$ in place of F_0 and F_1, and we write $d_{y,n} = \|\widehat{F}_y - F_y\|_\infty$ for the largest error in $\widehat{F}_y$.

Assumption 5 states the sampling conditions and the counterpart of Assumption 1(ii): on an interval $\mathcal{C}$ of thresholds, the worst-case FPR decreases at a rate bounded away from zero and the worst-case TPR decreases at a bounded rate.

Assumption 5 (Sampling and regularity under a pointwise budget)**.**

(i) Conditional on the independent training sample, the validation cases are independent draws from P_0, with $0 < p_0 < 1$, and the local extrema in Definition 6 are measurable.

(ii) For an FPR interval $I = [\alpha_l, \alpha_u] \subset (0,1)$, there are an interval $\mathcal{C} = [c_l, c_u] \subset (0,1)$ and $b > 0$ such that

$$1 - F_0(c_l) > \alpha_u + b, \qquad 1 - F_0(c_u) < \alpha_l - b,$$

[6] The distance of each proposal from the case is therefore uniform on $[0,\rho]$, so the proposals are not uniform over the ball.

F_0 and F_1 are continuously differentiable on a neighborhood of $\mathcal{C}$, and their derivatives $f_y = F_y'$ satisfy

$$0 < \underline{f} \le f_0(c), \qquad 0 \le f_1(c) \le L_1 < \infty \quad (c \in \mathcal{C}).$$

When the score takes finitely many values, as a finite tree ensemble does, F_0 and F_1 have finite supports. Let $s_1 < s_2 < \cdots$ be the points of the two supports, and let $c_1 < \cdots < c_M$ list 0, 1 and these points without repetition. For every threshold c, the indicators $\mathbb{I}(Z_y(X) > c)$ equal, with probability one, those at one of $c_1, \ldots, c_M$, so only these thresholds matter. Let

$$A_k = 1 - F_0(c_k), \qquad B_k = 1 - F_1(c_k), \qquad \mathcal{L} = \{A_k : 1 \le k \le M\}$$

be the worst-case FPR and TPR at c_k and the set of FPR levels that these thresholds reach. At FPR level α, $\tilde{g}_\rho(\alpha; P_0) = B_{k(\alpha)}$, where $k(\alpha) = \min\{k : A_k \le \alpha\}$ is the index of the smallest of these thresholds whose worst-case FPR is at most α. Hence $\tilde{g}_\rho(\cdot; P_0)$ is a step function that can jump only at levels in $\mathcal{L}$.

Theorem 5 (Estimation of the worst-case ROC curve under a pointwise budget)**.** *Suppose that Assumption 5(i) holds.*

(i) For continuous or discrete score distributions, $d_{0,n} + d_{1,n} = O_p(n^{-1/2})$. On samples with both classes, whenever $\alpha \pm d_{0,n} \in [0, 1]$,

$$\tilde{g}_\rho(\alpha - d_{0,n}; P_0) - d_{1,n} \le \widehat{\tilde{g}}_\rho(\alpha) \le \tilde{g}_\rho(\alpha + d_{0,n}; P_0) + d_{1,n}. \tag{4.13}$$

(ii) Under Assumption 5(ii),

$$\sup_{\alpha \in I} |\widehat{\tilde{g}}_\rho(\alpha) - \tilde{g}_\rho(\alpha; P_0)| = O_p(n^{-1/2}).$$

(iii) Suppose instead that F_0 and F_1 have fixed finite supports. Then the rate in part (ii) holds on every compact $I \subset (0, 1)$ with $\operatorname{dist}(I, \mathcal{L}) > 0$.

Unlike Theorem 1, Theorem 5(i) needs no complexity condition such as Assumption 1(i): the worst-case rates are one minus the CDFs of the worst-case scores, and empirical CDFs converge uniformly at the rate $n^{-1/2}$ for any F_0 and F_1. By (4.13), the estimated curve lies within $d_{0,n}$ horizontally and $d_{1,n}$ vertically of $\tilde{g}_\rho(\cdot; P_0)$. Where $\tilde{g}_\rho(\cdot; P_0)$ is steep, however, a horizontal shift of $d_{0,n}$ can change its value by much more than $d_{0,n}$, and where it jumps, the change can be as large as the jump. Parts (ii) and (iii) give two settings in which the estimated curve still converges at the rate $n^{-1/2}$.

In part (ii), let $c_\alpha = F_0^{-1}(1-\alpha)$ be the smallest threshold that is feasible at level α, so that $\tilde{g}_\rho(\alpha; P_0) = 1 - F_1(c_\alpha)$. By the endpoint inequalities in Assumption 5(ii), c_α lies in $\mathcal{C}$ for every α within b of I, and there the slope of the curve, $f_1(c_\alpha)/f_0(c_\alpha)$, is at most $L_1/\underline{f}$. By (4.13), once $d_{0,n} < b$, the error in the estimated curve on I is at most $(L_1/\underline{f})\, d_{0,n} + d_{1,n}$.

In part (iii), once $d_{0,n}$ is less than half the distance from I to $\mathcal{L}$, the estimated curve and $\tilde{g}_\rho(\cdot; P_0)$ use the same threshold $c_{k(\alpha)}$ at every $\alpha \in I$, so the error in the estimated curve on I is at most $d_{1,n}$. At a level in $\mathcal{L}$, an arbitrarily small error in $\widehat{F}_0$ can change the threshold that is used, and the estimated curve need not converge at that level (Appendix A.3), so finite supports alone do not give uniform convergence on every interval inside $(0, 1)$.

Under the conditions of Theorem 5(ii), the map from F_0 and F_1 to $\tilde{g}_\rho(\cdot; P_0)$ is also Hadamard differentiable, and under those of Theorem 5(iii), so is the map from the probabilities that F_0 and F_1 put on their support points (Theorem 7 in Appendix A.3).

Under a pointwise budget, the implemented rule is the rule of Section 3 with the following substitutions for the set, the worst-case rates and the worst-case ROC curve:

$$(\mathcal{B}_\varepsilon, g_\varepsilon, \overline{\alpha}_\varepsilon, \underline{\beta}_\varepsilon) \longmapsto (\tilde{\mathcal{B}}_\rho, \tilde{g}_\rho, \tilde{\alpha}_\rho, \tilde{\beta}_\rho).$$

It uses the same posteriors, set of candidate thresholds and maximizer in (3.4) as in Section 3.2. We say that a result of Section 3 holds with the substituted objects if it holds after these substitutions are made throughout. As in Section 3.2, the error in the estimated worst-case ROC curve includes any error from the grid of candidate thresholds.

Corollary 2 (Replacement and welfare under a pointwise budget)**.** *Suppose that* (2.4) *holds and that the measurable attaining selections assumed in Proposition 3 exist. All other hypotheses of the cited results, including their requirements on the candidate thresholds, are retained.*

(i) With Assumption 2 imposed on $\tilde{g}_\rho$ and under Assumption 3, Theorem 2 holds with the substituted objects.

(ii) Theorem 3 and, under Assumption 3, Corollary 1 hold with the substituted objects and the same $t_n(\gamma)$.

(iii) Theorem 4(i) holds with the substituted objects at every $Q \in \tilde{\mathcal{B}}_\rho(P_0)$ at which the benchmark is defined and Assumption 4 holds. If, in addition, Assumption 2(ii) is imposed on $\tilde{g}_\rho$ and Assumption 3 holds, so does Theorem 4(ii).

The welfare bounds of Corollary 1 and Theorem 4(ii) also hold for the welfare choice in (2.15)*.*

The logistic score. Differentiating (4.11) gives

$$\tilde{g}'_\rho(\alpha) = \frac{H_1'(a_\alpha + 2\rho)}{H_0'(a_\alpha)}.$$

Bounding H_1' from above at $a_\alpha + 2\rho$ and H_0' away from zero at a_α, for $\alpha \in [\alpha_l, \alpha_u]$, therefore gives Assumption 2(ii) for $\tilde{g}_\rho$ on $[\alpha_l, \alpha_u]$. Since $F_0(c) = H_0(a_c - \rho)$ and $F_1(c) = H_1(a_c + \rho)$ by (4.10), Assumption 5(ii) holds if its endpoint inequalities hold and, for every $c \in \mathcal{C}$, H_0 is continuously differentiable near $a_c - \rho$ with a derivative bounded away from zero and H_1 is continuously differentiable near $a_c + \rho$. Theorem 5(ii) then applies under Assumption 5(i). Although (4.9) gives the local extrema of the logistic score in closed form, the bounds in Corollary 2 still contain the errors in the estimated worst-case rates and the terms that come from the posteriors.

5 Monte Carlo simulations

We now compare the nominal and robust rules in simulations in which the data seen by the algorithm shift but the information used by the doctors does not. The distance from the training distribution to each deployment distribution is known, so we know, for each budget, which of them lie in the set. Section 5.1 describes the design. Section 5.2 compares the two rules at the prespecified average budgets, Section 5.3 turns to smaller average budgets, Section 5.4 to pointwise budgets, and Section 5.5 compares the two threshold choices.

5.1 Design

In past cases, let $Z \sim N(0, I_5)$ and $U \sim N(0, \sigma_u^2)$ be independent, and set

$$R = \xi_0 + Z^\top \xi + U, \qquad p(Z,U) = \{1 + \exp(-R)\}^{-1}, \qquad Y \mid Z, U \sim \text{Bernoulli}(p(Z,U)). \tag{5.1}$$

The algorithm sees $X = Z$, so the training distribution P_0 is the distribution of (Z, Y). The doctors also observe U, which stands for information that is not recorded, such as what a doctor learns by examining the patient. Doctor j decides according to

$$\widehat{Y}_j = \mathbb{I}\{p(Z,U) + \nu_j > c_{H,j}\}, \qquad \nu_j \sim N(0, \sigma_\nu^2), \qquad c_{H,j} = 0.15 + 0.7 V_j, \quad V_j \sim \text{Beta}(2,3).$$

Each doctor thus compares the probability that the patient is sick, observed with noise, with a threshold of the doctor's own. The noise makes the doctors' decisions less accurate than their information allows, so the algorithm can dominate some of them even though it does not see U. The thresholds lie between 0.15 and 0.85 and give the doctors different error rates. The noise is drawn independently for each case, and both the noise and the

thresholds are independent of (Z, U, Y). We set $\xi = (1, 0.5, -0.3, 0.2, -0.1)^\top$, $\xi_0 = -1.5$, $\sigma_u = 0.15$ and $\sigma_\nu = 0.225$.

At deployment, the algorithm sees

$$X_\delta = Z + (1 - 2Y)\delta e_1, \qquad \delta \geq 0, \tag{5.2}$$

instead of $X = Z$, where e_1 is the first coordinate vector. We write Q_δ for the deployment distribution, the distribution of (X_δ, Y), so that $Q_0 = P_0$. The distribution of (Z, U, Y, ν_j) and the doctors' decision rules do not change. Hence the prevalence does not change, and, by the sufficient condition of Section 2.1 with $\Omega_j = (Z, U, \nu_j)$, neither do the doctors' error rates, so (2.4) holds. The shift does change the conditional distribution of Y given X, so it is not a covariate shift.

The shift lowers the first coordinate of X by δ for every sick case and raises it by δ for every healthy case. Since $\xi_1 > 0$, sick cases tend to have larger values of this coordinate than healthy cases, so the shift moves each class toward the values that are typical of the other. The fitted logistic scores increase in this coordinate: in each of the 100 replications described below, its fitted coefficient lies between 0.93 and 1.08. As δ grows, the algorithm's TPR therefore falls and its FPR rises at every threshold $c \in (0, 1)$. One can think of the shift as a change in how one variable is recorded that moves sick and healthy cases in opposite directions.

Within each class, Q_δ is thus obtained from P_0 by moving every case by the same vector, of length δ. As in Section 2.2, let P^y and Q^y be the distributions of X given $Y = y$ under P_0 and Q_δ. Moving every case by this vector defines a coupling of P^y and Q^y with cost δ, so $W_1(P^y, Q^y) \leq \delta$. Conversely, $W_1(P^y, Q^y) \geq \delta$, because the means of P^y and Q^y differ by a vector of length δ. By (2.8), therefore,

$$\mathcal{W}_\kappa(P_0, Q_\delta) = \delta, \tag{5.3}$$

and we call δ the deployment distance. Hence $Q_\delta \in \mathcal{B}_\varepsilon(P_0)$ if and only if $\delta \leq \varepsilon$. These statements concern $\mathcal{B}_\varepsilon(P_0)$, the set to which the welfare guarantees of Section 3 refer. The rule itself uses the estimated worst-case rates, which are worst cases over $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_n)$.

We run 100 independent replications. Each replication draws, under (5.1), 10,000 training cases, 5,000 validation cases and, for each of $J = 500$ doctors, a history of 500 cases. The logistic score is fitted to the training cases by ℓ_2-penalized maximum likelihood and then held fixed.[7] The Dirichlet prior has parameter $a_0 = 0.01$ on each cell of the confusion matrix, and the cutoff is $\tau = 0.95$. For each average budget $\varepsilon \in \{0, 0.05, 0.1, 0.2, 0.3\}$, we compute the estimated worst-case rates exactly, as in Section 2.4, on a common grid of 1,001 thresholds from 0 to 1. For each doctor, we select the threshold on the grid that

[7] With the fitted score written as $m(x) = \{1 + \exp[-(w_0 + x^\top w)]\}^{-1}$, as in Appendix B, (w_0, w) minimizes $\sum_i \log[1 + \exp\{-(2Y_i - 1)(w_0 + X_i^\top w)\}] + \|w\|_2^2/2$ over the training cases.

maximizes (2.18) and replace the doctor if the maximum is at least τ. With $\varepsilon = 0$, this is the nominal rule, which replaces a doctor when (2.3) holds.

We evaluate the rules at the deployment distances $\delta \in \{0, 0.025, 0.05, 0.1, 0.3, 0.5, 1\}$. For the main budget $\varepsilon = 0.05$, $Q_{0.025}$ lies in the interior of $\mathcal{B}_{0.05}(P_0)$, and $Q_{0.05}$ lies on its boundary. Welfare (3.12) is evaluated with the prespecified weights $(\phi, \eta) \in \{(1,1), (2,1), (5,1)\}$, with $(2,1)$ as the main case. The replacement decisions and the selected thresholds depend on neither the deployment distribution nor the welfare weights, so, within a replication, they are the same at every deployment distance and for every pair of weights.

The error rates of the algorithm and of the doctors are computed by numerical integration under (5.1) and (5.2) rather than from a test sample (Appendix B), and the matched-FPR benchmark uses the same rates. For each rule and each deployment distance, we report the gain over retention, the share of doctors replaced and the share of harmful replacements, that is, the share of replaced doctors j for whom $W_M(\widehat{c}_j; Q_\delta) < W_{H,j}$. We also report the three sums in (3.15), which split the regret into the contributions of the common, missed and extra replacements, and the threshold term $\widehat{K}_n$ in (3.17), computed from the estimated worst-case ROC curve on the grid, the posterior medians $a_{n,j}$ and the selected thresholds. Monte Carlo standard errors are computed across replications, and differences between rules are paired within replications.

5.2 Replacement, welfare and regret

With any of the average budgets $\varepsilon \in \{0.05, 0.1, 0.2, 0.3\}$, the robust rule retains all 500 doctors in all 100 replications, whereas the nominal rule replaces 41.51% of them on average (Monte Carlo standard error 0.88 percentage points). Under the robust rule, the posterior probabilities of dominance are far below the cutoff: at $\varepsilon = 0.05$, the largest over all doctors and replications is 5.85×10^{-10}. Figure 3 shows why for the first replication. Every doctor's error rates lie below the estimated nominal ROC curve (panel (a)), although the nominal rule replaces only 159 of the 500 doctors, because a history of 500 cases does not establish dominance with probability 0.95 for doctors close to the curve. With $\varepsilon = 0.05$, the estimated worst-case ROC curve lies so far below the nominal curve that no doctor's error rates lie below it (panel (b)). Panel (c) shows the curve under a pointwise budget of the same size, to which we return in Section 5.4.

Table 1 compares the two rules at each deployment distance. The welfare of the robust rule is that of retaining all doctors, 0.6498, at every δ, because the doctors' error rates do not depend on δ. At $\delta = 0$, the nominal rule raises mean welfare from 0.6498 to 0.7323, a paired gain of 0.0826 (0.0017), and at $\delta = 0.05$, where Q_δ lies on the boundary of $\mathcal{B}_{0.05}(P_0)$, its gain over retention is still 0.0628 (0.0013). The mean gain over retention of the nominal rule is positive at every deployment distance up to $\delta = 0.1$ and negative

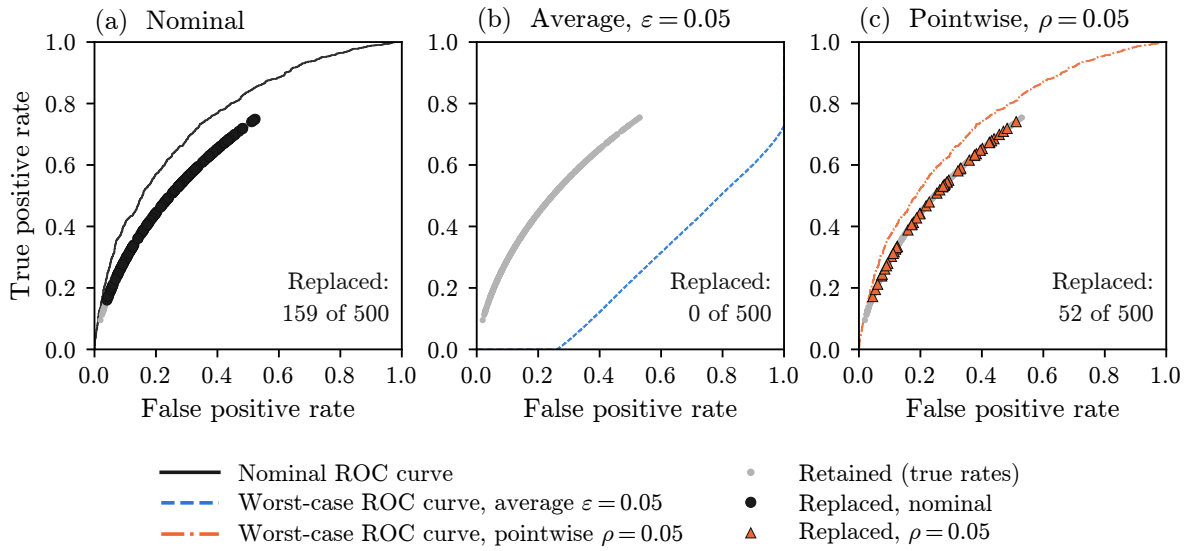


Figure 3: Estimated worst-case ROC curves and replaced doctors in the first replication.

Notes: The lines are the estimated worst-case ROC curves of the fitted logistic score in the first replication, computed from its validation sample: (a) the nominal rule; (b) the average budget $\varepsilon = 0.05$; (c) the pointwise budget $\rho = 0.05$. The points are the true error rates of the 500 doctors, and the highlighted points mark the doctors whom each rule replaces in this replication: 159, 0 and 52. The replacement decisions use the estimated worst-case rates from the validation sample and each doctor's history, rather than the plotted true human rates. A doctor below a curve therefore need not be replaced.

Table 1: Welfare of the nominal and robust rules by deployment distance, $(\phi, \eta) = (2, 1)$.

δ	Nominal welfare	Robust welfare	Robust minus nominal	Harmful among nominal replacements (%)
0	0.7323 (0.0020)	0.6498 (0.0009)	-0.0826 (0.0017)	0.17 (0.03)
0.025	0.7225 (0.0019)	0.6498 (0.0009)	-0.0727 (0.0015)	0.31 (0.05)
0.05	0.7126 (0.0017)	0.6498 (0.0009)	-0.0628 (0.0013)	0.70 (0.08)
0.1	0.6926 (0.0014)	0.6498 (0.0009)	-0.0428 (0.0009)	2.82 (0.20)
0.3	0.6105 (0.0013)	0.6498 (0.0009)	0.0392 (0.0010)	99.94 (0.02)
0.5	0.5270 (0.0027)	0.6498 (0.0009)	0.1228 (0.0027)	100.00 (0.00)
1	0.3272 (0.0067)	0.6498 (0.0009)	0.3226 (0.0068)	100.00 (0.00)

Notes: Means over 100 replications, with Monte Carlo standard errors in parentheses; welfare differences are paired within replications. The nominal rule has $\varepsilon = 0$, and the robust rule has $\varepsilon = 0.05$. The other prespecified average budgets, 0.1, 0.2 and 0.3, give the same decisions as 0.05 in every replication: no doctor is replaced. Robust welfare is therefore the welfare of retaining all doctors, and the share of harmful replacements is not defined for the robust rule. The distributions Q_δ with $\delta \leq 0.05$ lie in $\mathcal{B}_{0.05}(P_0)$. Error rates are computed by numerical integration, as in Appendix B.

at $\delta = 0.3$, 0.5 and 1. At $\delta = 1$, retaining all doctors is better by 0.3226 (0.0068). With these average budgets, the robust rule avoids the losses after large shifts only by retaining every doctor, so it also gives up the gains after small shifts.

Among the doctors replaced by the nominal rule, the mean share of harmful replacements is 0.17% at $\delta = 0$, 0.70% at $\delta = 0.05$ and 99.94% at $\delta = 0.3$, and every replacement is harmful at $\delta = 0.5$ and $\delta = 1$. The share does not measure the size of the losses: it is 100% at both $\delta = 0.5$ and $\delta = 1$, but the loss of the nominal rule relative to retention grows from 0.1228 to 0.3226. Under the other two pairs of welfare weights, the paired welfare differences between the two rules have the same signs as in Table 1 (Appendix B.1).

The regret can be split as in (3.15). At $\delta = 0$, the regret of the nominal rule is 0.17182 on average. Of this, 0.02857 comes from the common replacements, for which the benchmark and the nominal rule use different thresholds, and 0.14325 comes from the missed replacements. There are no extra replacements: in each of the 100 replications, every doctor lies below the algorithm's ROC curve under P_0, so at $\delta = 0$ the benchmark replaces every doctor, and the missed replacements are the doctors whom the nominal rule retains. In the bound (3.20), the choice of threshold enters through the threshold term $\widehat{K}_n$, which, unlike the first sum in (3.15), can be computed from the data. For the nominal rule, $\widehat{K}_n$ averages 0.0327 (0.0008). The robust rule replaces nobody, so its $\widehat{K}_n$ is zero, and its regret of 0.1864 at $\delta = 0.05$ comes entirely from missed replacements.

5.3 Smaller average budgets

With every positive average budget in Table 1, the robust rule retains all doctors. We therefore also use a finer grid of seven average budgets, $\varepsilon \in \{0, 0.0001, 0.0003, 0.001, 0.003, 0.01, 0.05\}$, and keep the samples and the evaluation unchanged. Figure 4 shows the gain over retention of the rule with each budget at each deployment distance.

As the budget grows, the estimated worst-case ROC curve falls quickly (Figure B.1), and the share of doctors replaced falls with it, from 22.79% at $\varepsilon = 0.0001$ to 0.14% at $\varepsilon = 0.003$; nobody is replaced at 0.01 or 0.05. At $\varepsilon = 0.0001$, the gain over retention is 0.0463 at $\delta = 0$, against 0.0826 for the nominal rule, and the loss is 0.0686 at $\delta = 0.5$, against 0.1228 for the nominal rule. Both the gains after small shifts and the losses after large ones shrink toward zero as the budget grows.

For all three pairs of welfare weights and at every deployment distance up to $\delta = 0.1$, the rules that replace some doctors, those with $\varepsilon = 0.0001$, 0.0003, 0.001 and 0.003, have lower mean welfare than the nominal rule but higher mean welfare than retaining all doctors. At $\delta = 0.3$, 0.5 and 1, the ranking reverses: on average, these rules lose less than the nominal rule, but they do worse than retaining all doctors. At $\delta = 0.5$ and 1, every replacement they make is harmful, in every replication with a replacement.

For these small budgets, the deployment distances above lie outside the set, except

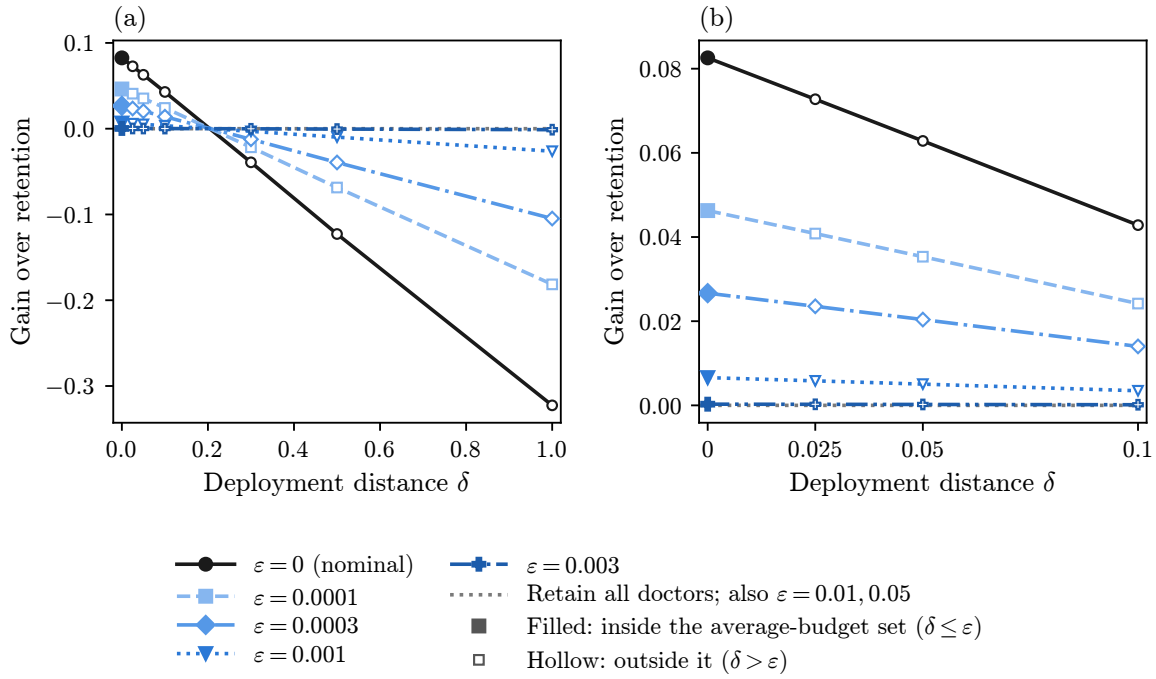


Figure 4: Gain over retention by average budget and deployment distance.

Notes: Mean gain over retention across the 100 replications, $(\phi, \eta) = (2, 1)$: (a) all seven deployment distances; (b) $\delta \leq 0.1$. Filled markers indicate the deployment distances at which Q_δ lies in $\mathcal{B}_\varepsilon(P_0)$, that is, $\delta \leq \varepsilon$, and hollow markers those outside it. For the budgets drawn as separate lines, $\delta = 0$ is the only such distance; Table 2 reports evaluations at $\delta = \varepsilon$. The rules with $\varepsilon = 0.01$ and 0.05 replace nobody and coincide with the zero line. Lines join the evaluated deployment distances. Table B.2 reports the values with Monte Carlo standard errors.

$\delta = 0$. We therefore also evaluate each rule at $\delta = \varepsilon/2$ and $\delta = \varepsilon$, where Q_δ lies inside the set and on its boundary, keeping its replacement decisions and thresholds. These distributions move every case by the same distance, so they are not the worst case in the set. Table 2 reports the results at $\delta = \varepsilon$. Each rule that replaces some doctors raises mean welfare there, by 0.0463 at $\varepsilon = 0.0001$ and by 0.0003 at $\varepsilon = 0.003$, and the results at $\delta = \varepsilon/2$ are almost the same. Because these budgets are small, so are the shifts inside their sets, and the gains are close to those at $\delta = 0$. The nominal rule gains more at the same deployment distances: the paired differences are negative for every budget and all three pairs of welfare weights (Table B.3 in Appendix B.2). At these deployment distances, the robust rule thus keeps part of the nominal rule's gain, and the part it gives up is the price of its smaller losses after the large shifts in Figure 4.

5.4 Pointwise budgets

We now compare average and pointwise budgets, keeping the 100 fitted logistic scores, the validation samples, the histories and the grid of 1,001 thresholds. The pointwise budgets are $\rho \in \{0, 0.05, 0.1, 0.2, 0.3\}$, the same values as the prespecified average budgets, and the deployment distances and welfare weights are as before. We compute the estimated

Table 2: Small average budgets at the boundary of their sets, $\delta = \varepsilon$.

	Robust rule			Nominal rule	
ε	Replaced (%)	Gain over retention	Harmful (%)	Gain over retention	Robust minus nominal
0.0001	22.79	0.04626	0.12	0.08253	−0.03627
	(0.70)	(0.00138)	(0.04)	(0.00173)	(0.00053)
0.0003	12.98	0.02662	0.11	0.08245	−0.05583
	(0.49)	(0.00099)	(0.06)	(0.00173)	(0.00089)
0.001	3.18	0.00656	0.00	0.08218	−0.07562
	(0.17)	(0.00035)	(0.00)	(0.00172)	(0.00145)
0.003	0.14	0.00028	0.00†	0.08139	−0.08112
	(0.02)	(0.00004)	(0.00)	(0.00171)	(0.00169)
0.01	0.00	0.00000	–	0.07865	−0.07865
	(0.00)	(0.00000)		(0.00165)	(0.00165)
0.05	0.00	0.00000	–	0.06284	−0.06284
	(0.00)	(0.00000)		(0.00133)	(0.00133)

Notes: Means over the 100 replications, with Monte Carlo standard errors in parentheses; welfare differences are paired within replications, and $(\phi, \eta) = (2, 1)$. Each rule keeps its replacement decisions and thresholds and is evaluated at $\delta = \varepsilon$, where Q_δ lies on the boundary of $\mathcal{B}_\varepsilon(P_0)$. The nominal rule replaces 41.51% of the doctors (0.88). The share of harmful replacements among the replaced doctors is averaged over the replications with replacements; † marks a mean over the 44 such replications, and with $\varepsilon = 0.01$ and 0.05, nobody is replaced. Table B.3 reports the other two pairs of welfare weights.

worst-case rates under a pointwise budget exactly, using (4.9). Because (5.2) moves every case by δ, Q_δ lies in $\tilde{\mathcal{B}}_\rho(P_0)$ if $\delta \leq \rho$. Conversely, $\tilde{\mathcal{B}}_\rho(P_0)$ is contained in $\mathcal{B}_\varepsilon(P_0)$ with $\varepsilon = \rho$ (Section 4.1), so $\delta \leq \rho$ by (5.3) whenever Q_δ lies in $\tilde{\mathcal{B}}_\rho(P_0)$. Hence Q_δ lies in $\tilde{\mathcal{B}}_\rho(P_0)$ if and only if $\delta \leq \rho$.

Average and pointwise budgets. With the prespecified positive average budgets, the robust rule replaces nobody (Section 5.2), but with a pointwise budget, the rule replaces 18.68%, 5.61% and 0.148% of the doctors at $\rho = 0.05$, 0.1 and 0.2, and nobody at $\rho = 0.3$. At $\rho = 0.2$, only 47 of the 100 replications have any replacement. The difference comes from how the two budgets move the worst-case rates at a fixed threshold: by an amount of order $\sqrt{\varepsilon}$ under an average budget and of order ρ under a pointwise budget (Remark 1 and Section 4.1). In the first replication, the estimated worst-case ROC curve with $\rho = 0.05$ stays close to the nominal one, and the rule replaces 52 of the 500 doctors (Figure 3(c)). For $(\phi, \eta) = (2, 1)$, the mean gain over retention is 0.0275 (0.0009) at $(\rho, \delta) = (0.05, 0.05)$ and 0.0052 (0.0003) at $(0.1, 0.1)$, in both cases with Q_δ on the boundary of $\tilde{\mathcal{B}}_\rho(P_0)$. With $\rho = 0.05$, 0.1 and 0.2, the rule raises mean welfare at every deployment distance at which Q_δ lies in $\tilde{\mathcal{B}}_\rho(P_0)$ and lowers it at $\delta = 0.3$, 0.5 and 1. After these large shifts, the rule with $\rho = 0.05$ loses about half as much as the nominal rule; at $\delta = 1$, for example, its loss is 0.1485, against 0.3226. Figure 5 shows the gains at all deployment distances, and

Table B.4 reports them with Monte Carlo standard errors.

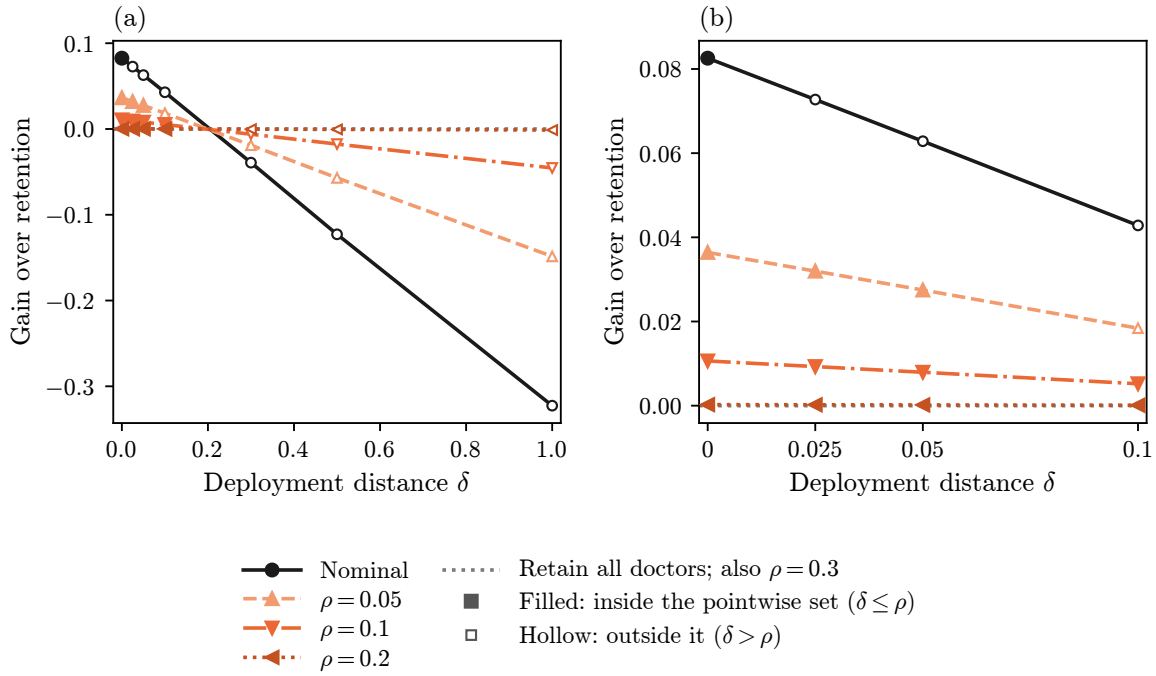


Figure 5: Gain over retention under a pointwise budget.

Notes: Mean gain over retention with the selected threshold $\widehat{c}^{\eta}$, $(\phi, \eta) = (2, 1)$, across the 100 replications: (a) all seven deployment distances; (b) $\delta \leq 0.1$. Filled markers indicate the deployment distances at which Q_δ lies in $\tilde{\mathcal{B}}_\rho(P_0)$, that is, $\delta \leq \rho$, and hollow markers those outside it. The curve for $\rho = 0$ is the nominal rule, which is the same under either budget. The rule with $\rho = 0.3$ and the rules with the prespecified positive average budgets coincide with the zero line. Table B.4 reports the means and Monte Carlo standard errors.

Because the rule uses estimated worst-case rates and posteriors from finite histories, individual replacements can be harmful even when Q_δ lies in $\tilde{\mathcal{B}}_\rho(P_0)$. The mean share of harmful replacements is 0.74% at $(\rho, \delta) = (0.05, 0.05)$ and 4.04% at $(0.1, 0.1)$ (Table B.5). At $(\rho, \delta) = (0.2, 0.1)$, it is 11.52% in the 47 replications with replacements. Because so few doctors are replaced, however, the harmful replacements make up only 0.016% of all doctors, averaged over all 100 replications. At $(\rho, \delta) = (0.2, 0.1)$, the gain over retention is negative in five of these 47 replications. At $\delta = 0.5$ and 1, every replacement is harmful.

The worst case in the set. The shift (5.2) lies in $\tilde{\mathcal{B}}_\rho(P_0)$ when $\delta \leq \rho$, but it is not the worst case in this set, because it moves the cases along the first coordinate rather than along w. For $0 \leq a \leq \rho$, let $T_a(x, y) = x + (1 - 2y)a\, w/\|w\|_2$, computed with the fitted coefficients of each replication, so that T_ρ is the map in (4.12). This map attains both worst-case rates under the pointwise budget at every threshold. By (2.4) and (4.8), the distribution of $(T_\rho(X, Y), Y)$ therefore gives each rule, with its replacement decisions and thresholds fixed, the lowest welfare over $\tilde{\mathcal{B}}_\rho(P_0)$. Figure 6 reports the rules under this distribution.

With $\rho = 0.05$ and 0.1, the rule still raises mean welfare, by 0.0259 (0.0009) and 0.0042 (0.0002). Under the same distributions, the nominal rule gains more, 0.0593 and 0.0356, but it loses 0.0126 at $\rho = 0.2$ and 0.0617 at $\rho = 0.3$. With $\rho = 0.2$, the rule replaces 0.148% of the doctors and lowers mean welfare by 0.000094 (0.000017); with $\rho = 0.3$, it replaces nobody. Under T_a with $a = \rho/2$, whose distributions lie inside the set, the gains are 0.03116, 0.00745 and 0.00008 for $\rho = 0.05$, 0.1 and 0.2 (Table B.6 in Appendix B.3).

At $\rho = 0.2$, 70 of the 74 replacements over the 100 replications are harmful; averaged over the 47 replications with replacements, the share is 94.68% (Figure 6(b)). These harmful replacements make up only 0.14% of all doctors, and the mean loss is small only because few doctors are replaced. The cutoff τ is a lower bound on each replaced doctor's posterior probability of dominance, computed with the estimated worst-case rates; it does not bound the share of harmful replacements. A replacement is harmful only if the doctor's true error rates lie outside the rectangle of the selected threshold, which has posterior probability at least τ, or if, at that threshold, the estimated worst-case rates are more favorable to the algorithm than the worst-case rates under P_0.

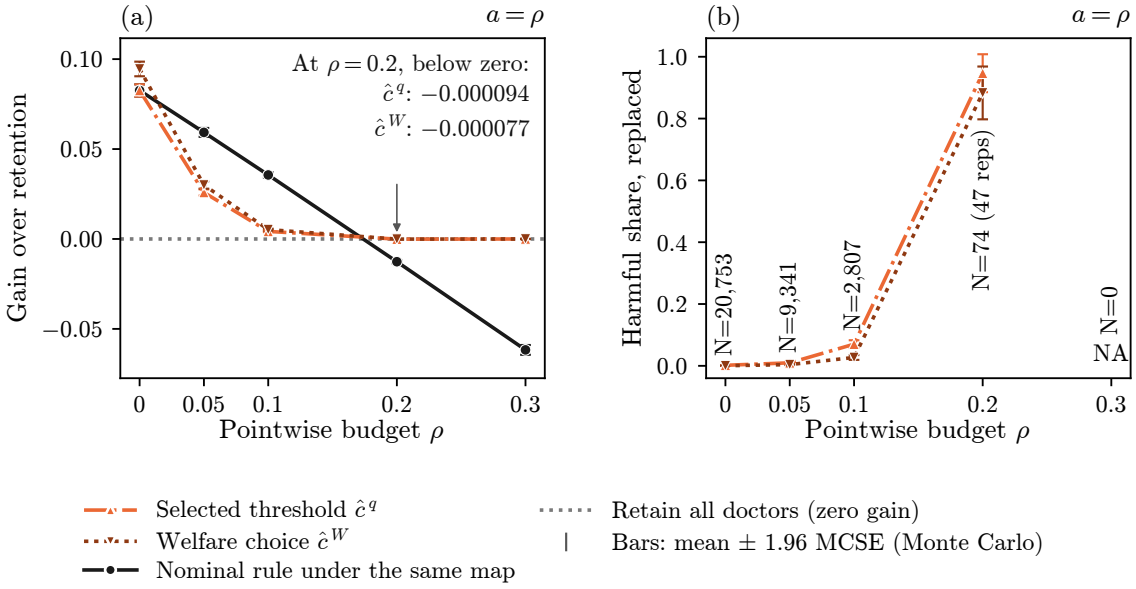


Figure 6: Pointwise rules under the worst case in their sets.

Notes: The rules of Figure 5, with both threshold choices, are evaluated under the distribution of $(T_\rho(X, Y), Y)$, with T_ρ in (4.12) computed with the fitted coefficients of each replication; this distribution attains both worst-case rates under the pointwise budget ρ at every threshold. The nominal rule is evaluated under the same distributions, and the doctors' error rates do not change. (a) Mean gain over retention, $(\phi, \eta) = (2, 1)$; at $\rho = 0.2$, the means are -0.000094 for $\widehat{c}^q$ and -0.000077 for $\widehat{c}^W$, with Monte Carlo standard errors 0.000017 and 0.000015. (b) Share of harmful replacements among the replaced doctors, computed in each replication with replacements and averaged over these replications with equal weights. N is the total number of replacements over the 100 replications; at $\rho = 0.2$, the 74 replacements occur in 47 replications, and at $\rho = 0.3$, nobody is replaced. Bars show ±1.96 Monte Carlo standard errors.

5.5 The two threshold choices

Finally, we compare the selected threshold $\widehat{c}^q$ with the welfare choice $\widehat{c}^W$ in (2.15), using the pointwise budgets of Section 5.4 and the weights $(2, 1)$. Figure 7 shows the two choices for one doctor in the first replication, with $\rho = 0.05$. Twenty-one thresholds pass the posterior test. Among them, the welfare lower bound is highest at the lowest, 0.322, which is the welfare choice; its posterior probability of dominance is 0.9503, just above τ. The selected threshold, 0.332, has a posterior probability of 0.9713 but a lower welfare lower bound, 0.8179 against 0.8461.

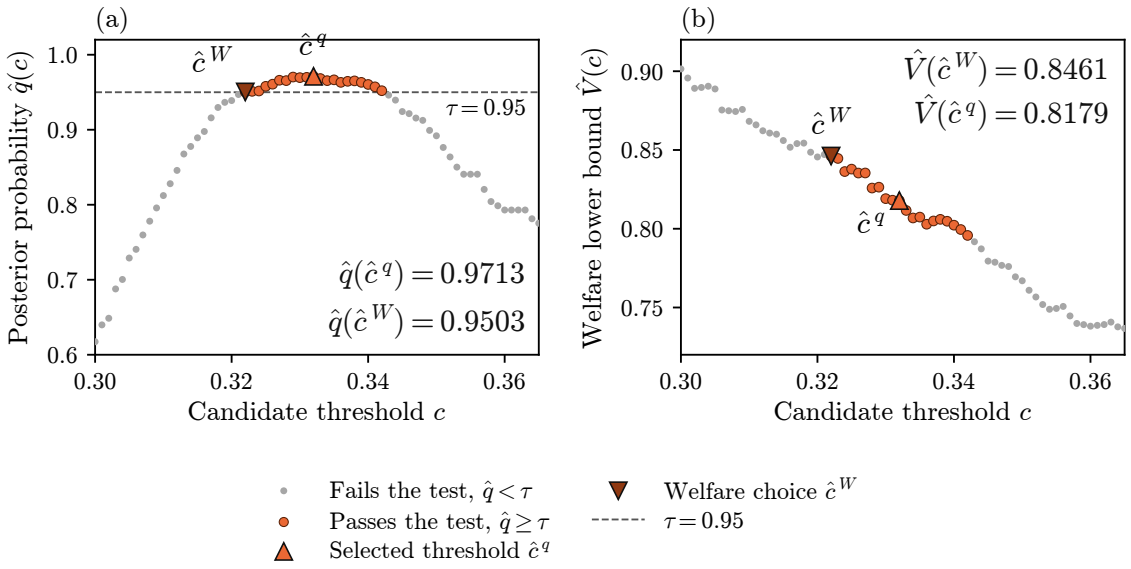


Figure 7: The selected threshold and the welfare choice for one doctor.

Notes: One doctor replaced in the first replication under the pointwise budget $\rho = 0.05$, with $(\phi, \eta) = (2, 1)$. Each dot is a candidate threshold between 0.30 and 0.365; this range contains 66 of the 1,001 candidates, including all 21 that pass the posterior test, which are drawn as larger dots with a dark edge. Here the thresholds that pass the test are consecutive, which need not hold in general. (a) Posterior probability of dominance, $\widehat{q}(c)$. (b) Welfare lower bound, $\widehat{V}(c) = \phi\tilde{\beta}_\rho(c; \widehat{\mathbb{P}}_n) - \eta\tilde{\alpha}_\rho(c; \widehat{\mathbb{P}}_n)$. The selected threshold $\widehat{c}^q$ maximizes $\widehat{q}$, and the welfare choice $\widehat{c}^W$ maximizes $\widehat{V}$ among the thresholds that pass the test.

Table 3 shows the same trade on average: the welfare choice gives up some posterior probability of dominance for a higher welfare lower bound. At $\rho = 0.05$, the mean posterior probability at the threshold used falls from 0.9774 with $\widehat{c}^q$ to 0.9564 with $\widehat{c}^W$, the welfare lower bound rises from 0.9003 to 0.9207, and the mean threshold term $\widehat{K}_n$ falls from 13.06×10^{-3} to 9.15×10^{-3}. In the paired comparison, welfare under $Q_{0.05}$ is higher by 4.15×10^{-3} (Monte Carlo standard error 0.18×10^{-3}), which raises the mean gain over retention to 0.0316.

The welfare choice has higher mean welfare at every deployment distance and for every pointwise budget with replacements, though not in every replication: at $(\rho, \delta) = (0.2, 1)$, seven replications show small negative differences. The same holds under the worst case in the set: the welfare choice raises the mean gain over retention from 0.0259 to 0.0300 at

Table 3: The two threshold choices under a pointwise budget.

ρ	Replaced (%)	Posterior probability $\widehat{c}^q$	Posterior probability $\widehat{c}^W$	Welfare lower bound $\widehat{c}^q$	Welfare lower bound $\widehat{c}^W$	$10^3\widehat{K}_n$ $\widehat{c}^q$	$10^3\widehat{K}_n$ $\widehat{c}^W$
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
0.3	0.00	–	–	–	–	0.00	0.00

Notes: Both choices use $\tau = 0.95$ and replace the same doctors; the welfare lower bound, $\widehat{K}_n$ and the welfare choice use the weights $(2, 1)$, and $\widehat{K}_n$ is computed with the worst-case rates under the pointwise budget, as in Section 4.2. Posterior probabilities and welfare lower bounds are averaged over the replaced doctors within a replication and then over the replications with replacements (100 at $\rho = 0$, 0.05 and 0.1; 47 at 0.2; none at 0.3). Shares replaced and $\widehat{K}_n$ are averaged over all 100 replications and 500 doctors. The row $\rho = 0$ is the nominal rule, which is the same under both budgets; the prespecified positive average budgets replace nobody. At $\rho = \delta = 0.05$, the paired difference in welfare under $Q_{0.05}$ between $\widehat{c}^W$ and $\widehat{c}^q$ is 4.15 (Monte Carlo standard error 0.18) in units of 10^{-3}; panel B of Table B.4 gives these differences for all budgets and deployment distances. Unrounded values are in the data supplement.

$\rho = 0.05$ and from 0.0042 to 0.0052 at $\rho = 0.1$, and it reduces the mean loss at $\rho = 0.2$ from 0.000094 to 0.000077, although 65 of its 74 replacements there are still harmful (Figure 6). As noted in Section 2.3, however, only the welfare choice requires the costs of the two errors.

6 An illustration with cross-state income data

We now apply the rules to income data from the 2018 American Community Survey (ACS). The ACSIncome task of Ding et al. (2021) is to predict, from ten demographic and employment variables, whether a person has high income, defined as annual income above \$50,000. The label Y equals one for high income. We pool the cases from three training states, California, New York and Texas; their distribution is the training distribution P_0. The evaluation states are five other states: West Virginia, Mississippi, New Mexico, North Dakota and Hawaii. In each of them, the distribution of cases plays the role of a deployment distribution.

We fit two scores, a logistic regression (LR) and a random forest (RF) with 100 trees, to 48,000 training cases. We use 6,000 further cases as the validation sample and another 6,000 as the histories of 200 synthetic decision makers, about 30 cases each. The decision makers stand in for the doctors. Their decisions take the form used in Section 5, with the LR's probability of high income in place of $p(Z, U)$. Each decision maker adds noise to this probability, once on the logit scale and once on the probability scale, and compares the result with a threshold of the decision maker's own. The two noise terms and the thresholds have the same distributions as U, ν_j and $c_{H,j}$ in Section 5. The noise is drawn independently of the labels, so, unlike U in Section 5, it carries no information

about income. Appendix C.1 describes how the cases are allocated, how the decisions are generated and how the rules are computed.

For the LR, the budget is an average budget, and we compute the estimated worst-case rates exactly, as in Section 2.4. For the RF, it is a pointwise budget, and we approximate the local extrema of the score by a search over $K = 200$ proposals, as in Section 4.1. Both budgets take the values 0, 0.1, 0.2, 0.3 and 0.5; with budget 0, the rule is the nominal rule. Distances are Euclidean, with each of the ten variables standardized by its mean and standard deviation in the training cases; the categorical variables enter as numeric codes (Appendix C.1). The prior and the cutoff are as in Section 5, with $a_0 = 0.01$ and $\tau = 0.95$. The replacement decisions and thresholds are computed from the cases of the training states only and are then applied in the evaluation states.

The nominal rule replaces 5 of the 200 decision makers with the LR and 10 with the RF. Each history has only about 30 cases, against 500 in Section 5. With every positive budget in this grid, the rule replaces nobody, whichever score is used. Even at the smallest positive budget, 0.1, the largest posterior probability of dominance among the 200 decision makers is 0.0374 with the LR and 0.8954 with the RF.

For the RF, the estimated worst-case ROC curve also depends on the number of proposals. At $\rho = 0.5$, we compare searches that use the first K proposals of one sequence of 1,000, for six values of K from 20 to 1,000. Using all 1,000 proposals instead of the first 200 lowers the area under the curve from 0.3937 to 0.2968, and nobody is replaced with any of the six values. A finite search can only understate the worst case (Section 4.1). At every threshold on the grid, the exact local extrema therefore give a worst-case FPR at least as high and a worst-case TPR at least as low as the search, and hence a posterior probability of dominance that is no higher. The rule computed with the exact local extrema would thus also replace nobody with any of the budgets 0.1, 0.2, 0.3 and 0.5, although we do not compute these extrema.

Table 4 reports welfare in the evaluation states. With either score, the nominal rule raises welfare in four of the five states and lowers it in North Dakota. The changes are small, between -0.0041 and 0.0191, because the nominal rule replaces only 5 or 10 decision makers. With the budgets from 0.1 to 0.5, the rule retains all decision makers, so it gives up the four gains and avoids the loss in North Dakota.

The evaluation states differ from the training states in more than the data seen by the algorithm. The share of cases with high income is 0.3942 in the validation sample and between 0.2695 and 0.3812 in the evaluation states (Table 4). Every distribution in $\mathcal{B}_\varepsilon(P)$ or $\tilde{\mathcal{B}}_\rho(P)$ has the prevalence of P (Sections 2.2 and 4.1), so the empirical distribution of each evaluation state lies outside these sets around the empirical distribution $\widehat{\mathbb{P}}_n$ of the validation sample, whatever the budget. The decision makers' decisions depend on the data seen by the algorithm, so the design does not impose (2.4), and their error rates differ across the evaluation samples. The welfare guarantees of Sections 3 and 4 therefore

Table 4: Welfare in the five evaluation states.

State	Prevalence	AUC		Welfare of retaining all	Gain over retention, nominal rule	
		LR	RF		LR	RF
WV	0.2906	0.8173	0.8404	1.0357	+0.00228	+0.00940
MS	0.2695	0.8225	0.8440	1.0703	+0.00614	+0.00767
NM	0.2977	0.8396	0.8659	1.0981	+0.00391	+0.01909
ND	0.3672	0.7996	0.8335	1.0442	-0.00407	-0.00377
HI	0.3812	0.8148	0.8491	1.0119	+0.00981	+0.01173

Notes: Welfare (3.12) with $(\phi, \eta) = (2, 1)$, that is, $2\text{TPR} - \text{FPR}$ averaged over the 200 decision makers with equal weights. AUC is the conventional area under the ROC curve in the evaluation state (Appendix C.1). Prevalence is the share of cases with high income; in the validation sample from the training states, it is 0.3942. The nominal rule replaces 5 decision makers with the LR and 10 with the RF. With the budgets from 0.1 to 0.5, nobody is replaced, so the gain over retention is zero in every state. All entries come from a single sample split and a single group of synthetic decision makers, not from replications.

do not apply to these states, and the results show only how rules computed from the training states perform in other states.

In a separate exercise, only the data seen by the algorithm change. We take 5,000 further California cases and add $\delta \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$ to the standardized age seen by the algorithm. The labels and the decision makers' decisions stay fixed, so the decision makers' error rates do not change with δ. Unlike in Section 5, δ is the distance $\mathcal{W}_\kappa$ between the distributions of the shifted and the unshifted California cases (Appendix C.1), not the distance from P_0. Both nominal rules have higher welfare than retaining all decision makers at every δ, whereas the rules with budget 0.5 retain all decision makers. Appendix C.2 reports the results for all budgets, numbers of proposals and age shifts.

Longer histories. The histories of the decision makers are short. To see how much this matters, we extend each decision maker's history, keeping its original cases, to 150 and then 300 cases, using further cases from the training states. The fitted scores, the validation sample and the decision makers stay the same. We also add smaller budgets and the LR with a pointwise budget. Table 5 reports how many decision makers each rule replaces. With 300 cases each, the nominal rule replaces 66 of the 200 decision makers with the LR and 139 with the RF, against 5 and 10 with the original histories. The budgets still matter: average budgets of 0.003 or more replace nobody at any history length, and so do pointwise budgets of 0.1 or more with the RF. With the LR and a pointwise budget, which moves the worst-case rates less than an average budget of the same size (Section 4.1), more decision makers are replaced: with 300 cases, 26 with a budget of 0.05 and 11 with a budget of 0.1. The short histories thus explain why the nominal rule replaces few decision makers, but not why the rules of the main analysis,

Table 5: Decision makers replaced in the income data, by budget and history length.

A. LR, average budget									
History	0	0.001	0.003	0.01	0.03	0.1	0.2	0.3	0.5
About 30 cases	5	1	0	0	0	0	0	0	0
150 cases	27	3	0	0	0	0	0	0	0
300 cases	66	9	0	0	0	0	0	0	0
B. LR, pointwise budget									
History	0	0.01	0.02	0.05	0.1	0.2	0.3	0.5	
About 30 cases	5	5	5	3	2	0	0	0	
150 cases	27	24	21	16	3	0	0	0	
300 cases	66	57	43	26	11	1	0	0	
C. RF, pointwise budget									
History	0	0.01	0.02	0.05	0.1	0.2	0.3	0.5	
About 30 cases	10	5	5	1	0	0	0	0	
150 cases	60	23	19	4	0	0	0	0	
300 cases	139	66	35	13	0	0	0	0	

Notes: Numbers of the 200 decision makers replaced. The original histories have 29 to 31 cases each. Each longer history contains the shorter ones, and the added cases come from the training states and are not used elsewhere. The fitted scores, the validation sample and the decision makers are the same in every row, and budget 0 gives the nominal rule, which is the same in panels A and B. In panel C, the local extrema of the RF are approximated by a search over $K = 200$ proposals, which can only understate the worst case (Section 4.1). Each history length is a single draw, and no standard errors are reported.

with budgets from 0.1 to 0.5, replace nobody. The data supplement reports the welfare of all these rules in the five evaluation states; the gains do not always rise with the history length.

7 Conclusion

We have proposed a rule for deciding whether an algorithm should replace a doctor when the deployment distribution is not known in advance. The rule replaces a doctor if, at some threshold, the posterior probability that the algorithm dominates the doctor under every distribution in a set around the training distribution reaches a cutoff. If the doctors' error rates do not change at deployment, the possible welfare loss relative to retaining all doctors is bounded by estimation errors that shrink as the validation sample and the histories grow.

The planner still has to choose the threshold and the budget. The threshold with the highest posterior probability of dominance need not have the highest welfare, even without estimation error. If the costs of the two errors can be specified, the welfare choice replaces the same doctors, and its threshold term is no larger. An average budget can be spent on the few cases closest to a threshold, so it moves the worst-case rates much

more than a pointwise budget of the same size. In our simulations, average budgets from 0.05 to 0.3 retain every doctor; of the average budgets we consider, only those of 0.003 or smaller replace some doctors. Pointwise budgets of 0.05 and 0.1 replace some doctors and raise mean welfare at every evaluated deployment distribution inside the set, including the worst case. A pointwise budget is justified only when each case can move by at most a known distance. How many doctors a rule can replace also depends on how much their histories reveal: in the income data, longer histories let the nominal rule replace many more decision makers, while average budgets of 0.003 or more still replace none.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used ChatGPT (OpenAI) to revise the wording, notation and organization of the manuscript, and Claude (Anthropic) to review drafts of the manuscript and the response to the referees and to suggest revisions to the text. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Adjaho, C. and Christensen, T. (2022). Externally valid policy choice. *arXiv preprint arXiv:2205.05561*. Version 4, November 2025.

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Annals of Statistics*, 35(2):608–633.

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

Chan, D. C., Gentzkow, M., and Yu, C. (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics*, 137(2):729–783.

Ding, F., Hardt, M., Miller, J. P., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490.

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 49(3):1378–1406.

Elliott, G. and Lieli, R. P. (2013). Predicting binary outcomes. *Journal of Econometrics*, 174(1):15–26.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

Fang, Z. and Santos, A. (2019). Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412.

Feng, K., Hong, H., and Nekipelov, D. (2026). Statistical inference of optimal allocations I: Regularities and their implications. *Journal of Econometrics*, 254:106217.

Feng, K., Hong, H., Tang, K., and Wang, J. (2025). Statistical tests for replacing human decision makers with algorithms. *Management Science*, 71(11):9145–9170.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., and Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286.

Firpo, S., Galvao, A. F., and Parker, T. (2023). Uniform inference for value functions. *Journal of Econometrics*, 235(2):1680–1699.

Gao, R. and Kleywegt, A. J. (2023). Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153.

Ho-Nguyen, N. and Wright, S. J. (2023). Adversarial classification via distributional robustness with Wasserstein ambiguity. *Mathematical Programming*, 198(2):1411–1447.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Kido, D. (2022). Distributionally robust policy learning with Wasserstein distance. *arXiv preprint arXiv:2205.04637*.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Kuhn, D., Mohajerin Esfahani, P., Nguyen, V. A., and Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Annals of Statistics*, 27(6):1808–1829.

Manski, C. F. (2011). Choosing treatment policies under ambiguity. *Annual Review of Economics*, 3:25–49.

Mo, W., Qi, Z., and Liu, Y. (2021). Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association*, 116(534):659–674.

Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166.

Mullainathan, S. and Obermeyer, Z. (2022). Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, 137(2):679–727.

Ötleş, E., Oh, J., Li, B., Bochinski, M., Joo, H., Ortwine, J., Shenoy, E., Washer, L., Young, V. B., Rao, K., and Wiens, J. (2021). Mind the performance gap: Examining dataset shift during prospective validation. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pages 506–534.

Shafieezadeh-Abadeh, S., Kuhn, D., and Mohajerin Esfahani, P. (2019). Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68.

Shafieezadeh-Abadeh, S., Mohajerin Esfahani, P., and Kuhn, D. (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, volume 28.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

Sugiyama, M. and Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

van der Vaart, A. W. and Wellner, J. A. (2011). A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203.

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., and Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683.

A Proofs and additional statistical results

A.1 Proofs for Section 2

The worst-case ROC curve. Because the algorithm uses a strict threshold, the event $\{m(X) > c\}$ is open when m is continuous, and the duality result of Blanchet and Murthy (2019) for closed events does not apply to it directly. For a nonempty set $A \subseteq \mathcal{X}$, write $d(x, A) := \inf_{a \in A} \|x - a\|_2$, and take closures relative to $\mathcal{X}$.

Lemma 1 (Worst-case probability of an event)**.** *Let $\mathcal{X}$ be a closed subset of $\mathbb{R}^d$, let μ have a finite first moment, and let A be a nonempty Borel subset of $\mathcal{X}$. For $r > 0$,*

$$\sup_{W_1(\nu,\mu) \le r} \nu(A) = \sup_{W_1(\nu,\mu) \le r} \nu(\overline{A}) \tag{A.1}$$

$$= \inf_{\lambda \ge 0} \left\{ \lambda r + \int (1 - \lambda d(x, A))_+ \, \mu(dx) \right\}. \tag{A.2}$$

Proof. We first prove (A.2) for a nonempty closed set F. The indicator $\mathbb{I}_F$ is bounded and upper semicontinuous, the Euclidean cost is nonnegative, continuous and zero only on the diagonal, and $\mathcal{X}$ is Polish. Theorem 1 of Blanchet and Murthy (2019), in the form for closed events of their Section 2.4, therefore applies. Its inner maximization is

$$\sup_{z \in \mathcal{X}} \{\mathbb{I}_F(z) - \lambda \|x - z\|_2\} = (1 - \lambda d(x, F))_+,$$

which gives (A.2) for F.

We next prove (A.1), for which $\mathbb{I}_A$ need not be upper semicontinuous. Put $F = \overline{A}$, and fix ν with $W_1(\nu, \mu) \le r$ and $t \in (0, 1)$. Mixing a transport plan from ν to μ with the identity plan shows that the mixture $\nu_t = (1 - t)\nu + t\mu$ satisfies

$$W_1(\nu_t, \mu) \le (1 - t) W_1(\nu, \mu) \le (1 - t) r.$$

Let $\{a_k : k \ge 1\}$ be a countable dense subset of A. For $\delta > 0$, let T_δ map each $x \in F$ to the first a_k with $\|x - a_k\|_2 < \delta$, and leave the points outside F unchanged. This map is Borel measurable, and $\|T_\delta(x) - x\|_2 < \delta$ on F. The distribution $\widetilde{\nu} = T_{\delta\#}\nu_t$ therefore satisfies

$$W_1(\widetilde{\nu}, \mu) \le \delta + (1 - t) r, \qquad \widetilde{\nu}(A) \ge \nu_t(F) \ge (1 - t)\nu(F).$$

With $0 < \delta < tr$, we have $W_1(\widetilde{\nu}, \mu) \le r$. Letting $t \downarrow 0$ and then taking the supremum over ν gives $\sup \nu(A) \ge \sup \nu(F)$, and the reverse inequality holds because $A \subseteq F$. Finally, $d(x, A) = d(x, F)$, so (A.2) for A follows from (A.2) for F. □

Proof of (2.8) and Proposition 1. We first prove (2.8). Under (2.6), a coupling with finite cost moves mass only within a class, so under both of its marginals the two classes have masses p and $1 - p$. Normalizing its two diagonal blocks gives a coupling of P^1 and Q^1 and one of P^0 and Q^0. Conversely, any two such couplings can be combined with weights p and $1 - p$. Taking infima over the two blocks proves (2.8). If the prevalences of P and Q differ, every coupling has infinite cost.

Write $S_c = \{m > c\}$ and $N_c = \mathcal{X} \setminus S_c = \{m \le c\}$. The TPR depends only on Q^1 and the FPR only on Q^0, so, for each rate, the whole budget can be spent on the class that

determines it, and (2.8) gives

$$\underline{\beta}_\varepsilon(c;P) = \inf_{W_1(\nu,P^1)\le\varepsilon/p} \nu(S_c),$$
$$\overline{\alpha}_\varepsilon(c;P) = \sup_{W_1(\nu,P^0)\le\varepsilon/(1-p)} \nu(S_c).$$

Because P has a finite first moment and $0 < p < 1$, P^1 and P^0 have finite first moments, so Lemma 1 applies with $\mu = P^1$ or $\mu = P^0$.

For the TPR, N_c is closed because m is continuous. Applying (A.2) to N_c and subtracting from one gives

$$\begin{aligned}\underline{\beta}_\varepsilon(c;P) &= 1 - \inf_{\lambda\ge0}\left\{\frac{\lambda\varepsilon}{p} + \int (1-\lambda d(x,N_c))_+ \, P^1(dx)\right\} \\ &= \sup_{\lambda\ge0}\left\{\int \min\{1,\lambda d_c^-(x)\}\, P^1(dx) - \frac{\lambda\varepsilon}{p}\right\}.\end{aligned}$$

For the FPR, S_c is open, and Lemma 1 with $r = \varepsilon/(1-p) > 0$ gives

$$\overline{\alpha}_\varepsilon(c;P) = \inf_{\lambda\ge0}\left\{\frac{\lambda\varepsilon}{1-p} + \int (1-\lambda d_c^+(x))_+ \, P^0(dx)\right\}.$$

Written as expectations under P, these two formulas are (2.10) and (2.11).

To see that λ can be restricted to $[0, 1/\varepsilon]$, multiply the two objectives by p and $1-p$, respectively, and write

$$T(\lambda) = \mathbb{E}_P[Y \min\{1,\lambda d_c^-(X)\}] - \lambda\varepsilon, \qquad F(\lambda) = \lambda\varepsilon + \mathbb{E}_P[(1-Y)(1-\lambda d_c^+(X))_+].$$

Then $T(0) = 0$, $T(\lambda) \le p - \lambda\varepsilon$, $F(0) = 1-p$ and $F(\lambda) \ge \lambda\varepsilon$. Hence no $\lambda > p/\varepsilon$ maximizes T, and no $\lambda > (1-p)/\varepsilon$ minimizes F. Both bounds are at most $1/\varepsilon$, so both optimizations can be restricted to $[0, 1/\varepsilon]$. The distances to the two nonempty decision regions are finite, so, by bounded convergence, both objectives are continuous on this interval, and their extrema are attained. The same bounds hold for the empirical distribution of any sample in which both classes occur. The statement of Proposition 1 for constant rules holds because $\{m(X) > c\}$ has probability zero under every distribution if S_c is empty, and probability one if $S_c = \mathcal{X}$. □

At $\varepsilon = 0$, the set contains only P, and the worst-case rates are the nominal rates. For nonempty S_c, the objective in (2.11) with $\varepsilon = 0$ is nonincreasing in λ, and monotone convergence as $\lambda \to \infty$ gives

$$\inf_{\lambda\ge0} \int (1-\lambda d_c^+(x))_+ \, P^0(dx) = P^0\{d_c^+(X) = 0\} = P^0(\overline{S}_c).$$

This can exceed the nominal FPR $P^0(S_c)$ (Section 2.2). For example, let $\mathcal{X} = \mathbb{R}$, $m(x) = (1+e^{-x})^{-1}$ and $c = 1/2$, and let P^0 be the point mass at zero. Because the threshold is strict, the nominal FPR is zero, whereas (2.11) with $\varepsilon = 0$ gives one. With any $\varepsilon > 0$, the worst-case FPR is one, because all the mass of the healthy cases can be moved a small distance to the right of zero.

The robust replacement decision.

Lemma 2 (Attainment of the worst-case ROC curve and equivalent posterior criteria)**.** *Let P be a distribution under which both classes have positive probability, let $\varepsilon \geq 0$, and let m be a measurable score with values in $[0,1]$; thresholds range over $[0,1]$. Let $A(c) = \overline{\alpha}_\varepsilon(c;P)$ and $B(c) = \underline{\beta}_\varepsilon(c;P)$ be the worst-case rates of Definition 1. For every $\alpha \in [0,1]$, the threshold*

$$c_\alpha := \min\{c \in [0,1] : A(c) \leq \alpha\}$$

exists and satisfies $g_\varepsilon(\alpha;P) = B(c_\alpha)$. For any posterior distribution Π_j on $[0,1]^2$,

$$\sup_{c\in[0,1]} \Pi_j\{\alpha_H \geq A(c),\ \beta_H \leq B(c) \mid \mathcal{D}_j\} = \sup_{\alpha\in[0,1]} \Pi_j\{\alpha_H \geq \alpha,\ \beta_H \leq g_\varepsilon(\alpha;P) \mid \mathcal{D}_j\}. \quad \text{(A.3)}$$

The conclusions also hold when P is the empirical distribution of a sample in which both classes occur, and for a finite grid of thresholds containing $c = 1$, with the worst-case ROC curve computed over that grid.

Proof. Both A and B are nonincreasing. If $c_n \downarrow c$, the events $\{m > c_n\}$ increase to $\{m > c\}$. By continuity from below under each distribution in the set, and after interchanging the two suprema,

$$\begin{aligned} A(c) &= \sup_{Q\in\mathcal{B}_\varepsilon(P)} \sup_n Q(m(X) > c_n \mid Y = 0) \\ &= \sup_n \sup_{Q\in\mathcal{B}_\varepsilon(P)} Q(m(X) > c_n \mid Y = 0) = \lim_n A(c_n), \end{aligned}$$

so A is right-continuous. Since $A(1) = 0$, the set of thresholds with $A(c) \leq \alpha$ is nonempty, and, because A is nonincreasing, it is an interval that extends to 1. By right-continuity it contains its infimum (when the infimum is 1, because $A(1) = 0$), so c_α exists. Every threshold that is feasible at level α satisfies $c \geq c_\alpha$, so $B(c) \leq B(c_\alpha)$ because B is nonincreasing, and therefore $g_\varepsilon(\alpha;P) = B(c_\alpha)$.

For (A.3), every threshold c is feasible at level $A(c)$, so $g_\varepsilon(A(c);P) \geq B(c)$. The event $\{\alpha_H \geq A(c), \beta_H \leq B(c)\}$ is therefore contained in $\{\alpha_H \geq A(c), \beta_H \leq g_\varepsilon(A(c);P)\}$, and the left side of (A.3) is at most the right side. Conversely, for every α, the threshold c_α satisfies $A(c_\alpha) \leq \alpha$ and $B(c_\alpha) = g_\varepsilon(\alpha;P)$, so

$$\{\alpha_H \geq \alpha,\ \beta_H \leq g_\varepsilon(\alpha;P)\} \subseteq \{\alpha_H \geq A(c_\alpha),\ \beta_H \leq B(c_\alpha)\}.$$

Taking posterior probabilities and suprema gives the reverse inequality. The same argument applies when P is an empirical distribution. On a finite grid that contains $c = 1$, the grid thresholds that are feasible at level α form a finite set, which is nonempty because $A(1) = 0$. Its smallest element is c_α, and the argument above, with the thresholds restricted to the grid, gives the other statements. □

If the supremum on the right side of (A.3) is attained at some α, the two inclusions in the proof show that c_α attains the supremum on the left side. In (A.3), the threshold is fixed before the posterior probability is taken, whereas in the event $\{\beta_H \leq g_\varepsilon(\alpha_H;P)\}$ the threshold may vary with (α_H, β_H).

Proof of Proposition 2. Fix P and the posterior. Since $P \in \mathcal{B}_\varepsilon(P)$, no worst-case rate is better than the corresponding rate under P, so $g_\varepsilon(\alpha;P) \leq g(\alpha;P)$ for every α. Hence, for each α,

$$\{\alpha_H \geq \alpha,\ \beta_H \leq g_\varepsilon(\alpha;P)\} \subseteq \{\alpha_H \geq \alpha,\ \beta_H \leq g(\alpha;P)\}.$$

Taking posterior probabilities at the same α and maximizing over α gives $\widehat{q}^{\varepsilon}_{\max,j} \le \widehat{q}^{0}_{\max,j}$. The two rules use the same cutoff τ, so every doctor replaced by the robust rule is also replaced by the nominal rule. With $\mathcal{B}_{\varepsilon'}(P) \subseteq \mathcal{B}_{\varepsilon}(P)$ in place of $P \in \mathcal{B}_{\varepsilon}(P)$, the same argument gives $\widehat{q}^{\varepsilon}_{\max,j} \le \widehat{q}^{\varepsilon'}_{\max,j}$ for $0 \le \varepsilon' \le \varepsilon$, so a doctor replaced at some budget is also replaced at every smaller budget. □

Computation.

Lemma 3 (Distances, complexity and computation for the logistic score)**.** *Let $m(x) = \{1 + \exp[-(w_0 + x^\top w)]\}^{-1}$ on $\mathbb{R}^d$, with $w \neq 0$, and restrict c to a compact subinterval of $(0,1)$. Then (2.19) holds. For fixed $\varepsilon > 0$ and $p_0 \in (0,1)$, the classes $\mathcal{F}^{\pm}$ are bounded, pointwise measurable VC-subgraph classes and satisfy Assumption 1(i).*

For an empirical distribution with $N_1, N_0 > 0$, put $n = N_1 + N_0$. For any nonnegative finite distances $d = (d_1, \dots, d_{N_y})$, define

$$V(d,\varepsilon) = \max\left\{ \frac{1}{n}\sum_{i=1}^{N_y} u_i : 0 \le u_i \le 1, \quad \frac{1}{n}\sum_{i=1}^{N_y} d_i u_i \le \varepsilon \right\}.$$

For $\varepsilon > 0$, the estimated worst-case rates are

$$\underline{\widehat{\beta}}_{\varepsilon}(c) = 1 - \frac{V((d_c^-(X_i))_{Y_i=1}, \varepsilon)}{N_1/n}, \qquad \widehat{\overline{\alpha}}_{\varepsilon}(c) = \frac{V((d_c^+(X_i))_{Y_i=0}, \varepsilon)}{N_0/n}.$$

The maximum that defines V is attained by setting $u_i = 1$ in increasing order of d_i until the budget is exhausted or all available mass has been moved, with a fractional value for the last case if necessary.

Proof. We prove the three claims in turn.

Distances. The algorithm classifies a case as positive when $z(x) > k_c$. If $z(x) > k_c$, the projection of x onto the hyperplane $\{z = k_c\}$ is $x - (z(x) - k_c)w/\|w\|_2^2$, at distance $(z(x) - k_c)/\|w\|_2$ from x, and by the Cauchy–Schwarz inequality no shorter move reaches $\{z \le k_c\}$. Points already in $\{z \le k_c\}$ are at distance zero. The distance to $\{z > k_c\}$ is found in the same way, by projecting in the opposite direction and, if needed, moving an arbitrarily small further distance into this open set. This proves both formulas in (2.19).

Complexity. Let $a = \lambda/\|w\|_2$ and $\ell(v) = \min\{1, \max\{0, v\}\}$. Before division by p_0 and $1 - p_0$, the functions in $\mathcal{F}^+$ and $\mathcal{F}^-$ are

$$y\ell\{az(x) - ak_c\}, \qquad (1-y)\ell\{1 + az(x) - ak_c\}.$$

Each is zero for one value of y and equals $\ell(az - ak_c)$ or $\ell(az + 1 - ak_c)$ for the other. For $0 \le u < 1$, the condition $u < \ell(v)$ is equivalent to $u < v$; for $u < 0$ it always holds, and for $u \ge 1$ it never does. On $0 \le u < 1$, each subgraph is therefore a half-plane in the coordinates (z, u), so each of the two classes of functions $\ell(az - ak_c)$ and $\ell(az + 1 - ak_c)$, indexed by (λ, c), has finite VC dimension. Multiplying by y or $1 - y$ sets a function to zero for the other value of y, and dividing by p_0 or $1 - p_0$ rescales u; neither operation changes the finiteness of the VC dimension. For each of the two classes $\mathcal{F}$, whose envelope is the constant $M = 1/p_0$ or $M = 1/(1-p_0)$, Theorem 2.6.7 of van der Vaart and Wellner (1996, p. 141), with $r = 2$, gives

$$N(uM, \mathcal{F}, L_2(H)) \le K_0 V(\mathcal{F})(16e)^{V(\mathcal{F})} u^{-2\{V(\mathcal{F})-1\}}, \qquad 0 < u < 1.$$

Here $V(\mathcal{F})$ is the VC index of the subgraphs, which is one more than their VC dimension, and K_0 is a universal constant. The bound holds for every probability distribution H. The functions are continuous in (λ, c) at every (x, y), and the parameter set is compact, so the functions indexed by a countable dense set of parameters approximate every member of the class pointwise. Hence the classes are pointwise measurable, and Assumption 1(i) holds.

Computation. The linear program that defines V is feasible and bounded. By linear programming duality, after the multipliers of the constraints $u_i \leq 1$ are eliminated,

$$V(d, \varepsilon) = \min_{\lambda \geq 0} \left\{ \lambda\varepsilon + \frac{1}{n} \sum_{i=1}^{N_y} (1 - \lambda d_i)_+ \right\}.$$

Substituting $\min\{1, v\} = 1 - (1 - v)_+$ into the dual for the TPR, and using the dual for the FPR as it stands, gives the two formulas for the estimated worst-case rates. In the primal problem, moving the mass of case i costs d_i per unit. Spending budget on a case with a smaller d_i instead of one with a larger positive d_i cannot reduce the mass moved, so, as in the fractional knapsack problem, some optimum moves the cases in increasing order of d_i, starting with those at distance zero. For the FPR, moving a case by d_i brings it only to the boundary of S_c, but by (A.1) the same value is obtained for the open event S_c. The dual objective is piecewise linear in λ, with knots at $1/d_i$ for positive d_i, and by Proposition 1 it can be minimized over $[0, 1/\varepsilon]$. It therefore suffices to evaluate it at 0, at $1/\varepsilon$ and at the knots in between, as in Section 2.4. □

A.2 Proofs for Section 3

Estimating the worst-case ROC curve.

Proof of Theorem 1. Write $A(c) = \overline{\alpha}_\varepsilon(c; P_0)$ and $B(c) = \underline{\beta}_\varepsilon(c; P_0)$, and write $\widehat{A}$ and $\widehat{B}$ for the estimated worst-case rates. The suprema over c in the error terms below are taken over $\mathcal{C}$, and we condition throughout on the independent training sample.

Step 1: uniform convergence of the dual objectives. By Proposition 1, both dual optimizations, under P_0 and under $\widehat{\mathbb{P}}_n$, can be restricted to $\Lambda = [0, 1/\varepsilon]$. Under the covering bound in Assumption 1(i), with constant envelope M, the uniform entropy integral is bounded by

$$\int_0^1 \sqrt{1 + v \log(K/u)}\, du < \infty.$$

The maximal inequality of van der Vaart and Wellner (2011, equation (1.1)) requires measurability and a bounded envelope, which Assumption 1(i) provides. The entropy integral is also finite for the union of the two classes, because the covering number of a union is at most the sum of the covering numbers of its parts. The inequality therefore gives

$$\sup_{f \in \mathcal{F}^+ \cup \mathcal{F}^-} |(\widehat{P}_n - P_0) f| = O_p(n^{-1/2}).$$

The estimated dual objectives divide by $\widehat{p}$ and $1 - \widehat{p}$ instead of p_0 and $1 - p_0$. Because $\lambda\varepsilon \leq 1$ on Λ, the numerator of the estimated TPR objective, $\widehat{\mathbb{P}}_n[Y \cdot \min\{1, \lambda d_c^-\}] - \lambda\varepsilon$, is at most one in absolute value, and that of the estimated FPR objective is at most two. Since $\widehat{p} - p_0 = O_p(n^{-1/2})$ and $0 < p_0 < 1$, $\widehat{p}$ is bounded away from zero and one with probability tending to one. On this event, replacing $\widehat{p}$ and $1 - \widehat{p}$ by p_0 and $1 - p_0$

changes each estimated objective by $O_p(n^{-1/2})$, uniformly in (λ, c). After this replacement, the estimated objectives differ from those under P_0 by $(\widehat{\mathbb{P}}_n - P_0)f$ with $f \in \mathcal{F}^+ \cup \mathcal{F}^-$. Hence both estimated dual objectives are within $O_p(n^{-1/2})$ of the dual objectives under P_0, uniformly on $\Lambda \times \mathcal{C}$. Samples in which one class is absent have exponentially small probability, and on them the estimated worst-case rates may be set to any bounded values without affecting the conclusion.

Step 2: the worst-case rates. Let

$$\delta_{\alpha,n} = \sup_c |\widehat{A}(c) - A(c)|, \qquad \delta_{\beta,n} = \sup_c |\widehat{B}(c) - B(c)|.$$

The maps $f \mapsto \sup_\lambda f$ and $f \mapsto \inf_\lambda f$ are 1-Lipschitz in the supremum norm, so Step 1 gives $\delta_{\alpha,n} + \delta_{\beta,n} = O_p(n^{-1/2})$, which is part (i).

Step 3: the inverse of the worst-case FPR. Suppose that $\delta_{\alpha,n} < b_C$. For $a \in [\alpha_l - b_C, \alpha_u + b_C]$, let c_a be the smallest threshold in $\mathcal{C}$ with $A(c) \le a$, as in the definition of c_α before Theorem 1. By the endpoint inequalities in Assumption 1(ii), $A(c_l) > a > A(c_u)$, so c_a is also the smallest threshold in $[0,1]$ that is feasible at level a, and $g_\varepsilon(a; P_0) = B(c_a)$ by Lemma 2. For $\alpha \in [\alpha_l, \alpha_u]$, let $\widehat{c}_\alpha$ be the smallest threshold in $[0,1]$ that is feasible at level α under $\widehat{\mathbb{P}}_n$, that is, with $\widehat{A}(c) \le \alpha$. By Lemma 2, $\widehat{c}_\alpha$ exists and $\widehat{g}_\varepsilon(\alpha) = \widehat{B}(\widehat{c}_\alpha)$, even if the estimated worst-case rates jump. Because $\delta_{\alpha,n} < b_C$, the endpoint inequalities also give $\widehat{A}(c_l) > \alpha_u$ and $\widehat{A}(c_u) < \alpha_l$, so $\widehat{c}_\alpha$ lies in $\mathcal{C}$.

If $c \in \mathcal{C}$ and $c < c_{\alpha+\delta_{\alpha,n}}$, then $A(c) > \alpha + \delta_{\alpha,n}$ and hence $\widehat{A}(c) > \alpha$. On the other hand, $\widehat{A}(c_{\alpha-\delta_{\alpha,n}}) \le A(c_{\alpha-\delta_{\alpha,n}}) + \delta_{\alpha,n} \le \alpha$. Therefore

$$c_{\alpha+\delta_{\alpha,n}} \le \widehat{c}_\alpha \le c_{\alpha-\delta_{\alpha,n}}. \tag{A.4}$$

Since $-A' \ge \underline{L}_\alpha$ on $\mathcal{C}$, the map $a \mapsto c_a$ is Lipschitz with constant $1/\underline{L}_\alpha$, and so

$$\sup_{\alpha \in [\alpha_l, \alpha_u]} |\widehat{c}_\alpha - c_\alpha| \le \delta_{\alpha,n}/\underline{L}_\alpha.$$

Step 4: the worst-case ROC curve. By the definition of $\delta_{\alpha,n}$, every threshold in $\mathcal{C}$ that is feasible at level $\alpha - \delta_{\alpha,n}$ under P_0 is feasible at level α under $\widehat{\mathbb{P}}_n$, and every threshold in $\mathcal{C}$ that is feasible at level α under $\widehat{\mathbb{P}}_n$ is feasible at level $\alpha + \delta_{\alpha,n}$ under P_0. Since $|\widehat{B} - B| \le \delta_{\beta,n}$ on $\mathcal{C}$, these two inclusions, applied to $c_{\alpha-\delta_{\alpha,n}}$ and to $\widehat{c}_\alpha$, which lies in $\mathcal{C}$ by Step 3, give

$$g_\varepsilon(\alpha - \delta_{\alpha,n}; P_0) - \delta_{\beta,n} \le \widehat{g}_\varepsilon(\alpha) \le g_\varepsilon(\alpha + \delta_{\alpha,n}; P_0) + \delta_{\beta,n}. \tag{A.5}$$

On $[\alpha_l - b_C, \alpha_u + b_C]$, the map $a \mapsto c_a$ is Lipschitz with constant $1/\underline{L}_\alpha$, and $|B'| \le L_\beta$ on $\mathcal{C}$, so $a \mapsto g_\varepsilon(a; P_0) = B(c_a)$ is Lipschitz with constant at most $L_\beta/\underline{L}_\alpha$. Hence

$$\sup_{\alpha \in [\alpha_l, \alpha_u]} |\widehat{g}_\varepsilon(\alpha) - g_\varepsilon(\alpha; P_0)| \le \delta_{\beta,n} + \frac{L_\beta}{\underline{L}_\alpha}\delta_{\alpha,n}.$$

Together with part (i) and $\mathbb{P}(\delta_{\alpha,n} < b_C) \to 1$, this proves part (ii). □

Theorem 1 assumes that the estimated worst-case rates are computed exactly and that all thresholds are searched. Otherwise, the uniform numerical error in these rates is added to the bounds, and on a finite grid so is the error from computing the estimated worst-case ROC curve over the grid points instead of over all thresholds (Section 3.2).

The limit distribution of the estimated worst-case ROC curve. Theorem 1 bounds the estimation error in the worst-case ROC curve. A limit distribution for the estimated curve, as a function of the FPR level, also requires a first-order expansion that is uniform in this level, and pointwise differentiability of an optimized value does not give such an expansion (Firpo et al., 2023). We give a sufficient condition in terms of the sets of optimal multipliers and verify it for the logistic score under the conditions of Remark 3 below.

Let $\mathcal{K} = \Lambda \times \mathcal{C}$ and

$$f^+_{\lambda,c}(x,y) = y \min\{1, \lambda d^-_c(x)\}, \qquad f^-_{\lambda,c}(x,y) = (1-y)(1-\lambda d^+_c(x))_+.$$

These are the dual functions $\psi^{\pm}_{\lambda,c}$ of Section 3.1 multiplied by p_0 and $1-p_0$, respectively. We represent a distribution P by the moments

$$\mathcal{S}(P) = \big(PY,\ (Pf^+_{\lambda,c})_{\mathcal{K}},\ (Pf^-_{\lambda,c})_{\mathcal{K}}\big) \in \mathbb{D} := \mathbb{R} \times \ell^\infty(\mathcal{K})^2, \tag{A.6}$$

where $\mathbb{D}$ is normed by the maximum of the norms of the three coordinates. We regard the worst-case ROC curve $g_\varepsilon(\cdot; P)$ on $[\alpha_l, \alpha_u]$ as a function of $\mathcal{S}(P)$. The map $\mathcal{S}(P) \mapsto g_\varepsilon(\cdot; P)$ is defined on the arrays $\mathcal{S}(P)$ of distributions P that satisfy the conditions of Proposition 1 and the endpoint inequalities $\overline{\alpha}_\varepsilon(c_l; P) > \alpha_u$ and $\overline{\alpha}_\varepsilon(c_u; P) < \alpha_l$. For such P, the smallest threshold that is feasible at a level in $[\alpha_l, \alpha_u]$ lies in $\mathcal{C}$, so, by Proposition 1 and Lemma 2, the curve on $[\alpha_l, \alpha_u]$ depends on P only through $\mathcal{S}(P)$. Directions of differentiation belong to $\mathbb{T} = \mathbb{R} \times C(\mathcal{K})^2$, where $C(\mathcal{K})$ is the space of continuous functions on $\mathcal{K}$. We call a family P_t of such distributions admissible if $\{\mathcal{S}(P_t) - \mathcal{S}(P_0)\}/t \to h \in \mathbb{T}$ in the norm of $\mathbb{D}$ as $t \downarrow 0$; the difference quotients themselves need not belong to $\mathbb{T}$. For example, if $P_t = P_0 + tH$ for a signed measure H, every difference quotient equals $(HY, (Hf^+_{\lambda,c})_{\mathcal{K}}, (Hf^-_{\lambda,c})_{\mathcal{K}})$, and when this array belongs to $\mathbb{T}$, it is the direction h. The sets of optimal multipliers under P_0 are

$$\Gamma_+(c) = \arg\max_{\lambda \in \Lambda} \{P_0 f^+_{\lambda,c} - \lambda\varepsilon\}, \qquad \Gamma_-(c) = \arg\min_{\lambda \in \Lambda} \{P_0 f^-_{\lambda,c} + \lambda\varepsilon\}.$$

For nonempty compact sets $\Gamma_1, \Gamma_2 \subset \Lambda$, the Hausdorff distance is

$$d_{\mathrm{Haus}}(\Gamma_1, \Gamma_2) = \max\Big\{ \sup_{\lambda \in \Gamma_1} \inf_{\lambda' \in \Gamma_2} |\lambda - \lambda'|,\ \sup_{\lambda' \in \Gamma_2} \inf_{\lambda \in \Gamma_1} |\lambda - \lambda'| \Big\}.$$

Finally, let $A_0(c) = \overline{\alpha}_\varepsilon(c; P_0)$ and $B_0(c) = \underline{\beta}_\varepsilon(c; P_0)$.

Theorem 6 (Directional differentiability of the worst-case ROC curve)**.** *Suppose that Assumption 1 holds, that $(\lambda, c) \mapsto f^{\pm}_{\lambda,c}$ is continuous into $L_2(P_0)$ on $\mathcal{K}$, and that $c \mapsto \Gamma_\pm(c)$ is continuous in the Hausdorff distance between nonempty compact subsets of Λ. Then the worst-case ROC curve, as a function of the moments (A.6), is Hadamard directionally differentiable at $\mathcal{S}(P_0)$ tangentially to $\mathbb{T}$, with values in $\ell^\infty([\alpha_l, \alpha_u])$. For $h = (h_p, h_+, h_-)$, the derivatives of the two worst-case rates are*

$$\dot{B}_h(c) = \frac{\max_{\lambda \in \Gamma_+(c)} h_+(\lambda, c) - B_0(c) h_p}{p_0}, \tag{A.7}$$

$$\dot{A}_h(c) = \frac{\min_{\lambda \in \Gamma_-(c)} h_-(\lambda, c) + A_0(c) h_p}{1 - p_0}. \tag{A.8}$$

They are continuous in c. The derivatives of the threshold c_α and of the worst-case ROC

curve are, uniformly in $\alpha \in [\alpha_l, \alpha_u]$,

$$\dot{c}_h(\alpha) = -\frac{\dot{A}_h(c_\alpha)}{A_0'(c_\alpha)}, \qquad \dot{g}_h(\alpha) = \dot{B}_h(c_\alpha) - \frac{B_0'(c_\alpha)}{A_0'(c_\alpha)}\dot{A}_h(c_\alpha). \tag{A.9}$$

In particular, for every admissible family P_t,

$$\left\| \frac{g_\varepsilon(\cdot; P_t) - g_\varepsilon(\cdot; P_0)}{t} - \dot{g}_h \right\|_\infty \longrightarrow 0.$$

If each set $\Gamma_\pm(c)$ *is a singleton, the derivative is linear and the map is fully Hadamard differentiable tangentially to* $\mathbb{T}$.

The condition on $\Gamma_\pm$ allows the optimal multipliers to be nonunique, provided that the sets vary continuously with c. It ensures that multipliers that nearly attain the optimum are uniformly close to the optimal sets, so that the expansion of the worst-case rates is uniform in c (Lemma 4 below).

Remark 3 (Primitive conditions for the logistic score)**.** *For the logistic score on* $\mathbb{R}^d$ *with* $w \neq 0$, *let* $S = z(X)/\|w\|_2$ *and* $a(c) = a_c = k_c/\|w\|_2$ *be as in Remark 1, so that a case is classified as positive when* $S > a(c)$. *Suppose that* $S \mid Y = y$ *has a continuous, strictly positive density on* $\mathbb{R}$ *and a finite first moment, and that, on a compact interval* $\mathcal{C}$ *of thresholds,*

$$\frac{\varepsilon}{p_0} < \mathbb{E}[(S - a(c))_+ \mid Y = 1], \qquad \frac{\varepsilon}{1 - p_0} < \mathbb{E}[(a(c) - S)_+ \mid Y = 0] \quad \text{for every } c \in \mathcal{C}. \tag{A.10}$$

Condition (A.10) *says that, at every threshold in* $\mathcal{C}$, *the budget is too small to move all the sick cases above the threshold onto it, so the worst-case TPR is positive. The budget is also too small to move all the healthy cases below the threshold across it, so the worst-case FPR is below one. Lemma 5 shows that the optimal multipliers are then unique and continuous in* c, *that the functions* $f^{\pm}_{\lambda,c}$ *are continuous in* $L_2(P_0)$, *and that both worst-case rates are continuously differentiable on* $\mathcal{C}$ *with strictly negative derivatives bounded away from zero. Since Lemma 3 gives Assumption 1(i), Theorem 6 applies when, in addition,* X *has a finite first moment, the validation cases are independent draws from* P_0, *and the FPR interval is chosen strictly inside* $A_0(\mathcal{C})$, *so that the endpoint inequalities in Assumption 1(ii) hold.*

Under the conditions of Theorem 6, the empirical process indexed by Y and the two classes $\{f^{\pm}_{\lambda,c}\}$ converges weakly in $\mathbb{D}$ to a tight Gaussian process $\mathbb{G}$ with paths in $\mathbb{T}$ and covariance function $P_0(fg) - (P_0 f)(P_0 g)$. The delta method for directionally differentiable maps (Fang and Santos, 2019, Theorem 2.1) then gives

$$\sqrt{n}\{\widehat{g}_\varepsilon - g_\varepsilon(\cdot; P_0)\} \rightsquigarrow \dot{g}_{\mathbb{G}} \quad \text{in } \ell^\infty([\alpha_l, \alpha_u]). \tag{A.11}$$

We prove (A.11) after the proof of Theorem 6. As in Theorem 1, the limit is conditional on the independent training sample, with the fitted score and the budget held fixed as the validation sample grows. It applies to the estimated worst-case ROC curve computed over all thresholds, with the estimated worst-case rates computed exactly. A curve computed over a grid or from numerically approximated rates has the same limit if its uniform difference from this curve is $o_p(n^{-1/2})$.

If the optimal multipliers are unique, the limit in (A.11) is Gaussian. Otherwise the directional derivative can be nonlinear, the limit need not be Gaussian, and (A.11) alone

does not justify the standard bootstrap (Fang and Santos, 2019, Section 3). By the chain rule, (A.11) also gives the limit distribution of smooth finite-dimensional functionals of the estimated worst-case ROC curve.

The proof of Theorem 6 uses the following lemma, which expands an optimized value uniformly in a parameter.

Lemma 4 (Uniform expansion of an optimized value)**.** *Let F be continuous on the product of compact metric spaces $\Lambda \times \mathcal{C}$, and let $V(c) = \max_{\lambda\in\Lambda} F(\lambda, c)$. Suppose that $c \mapsto \Gamma(c) = \arg\max_\lambda F(\lambda, c)$ is continuous in the Hausdorff distance. If $H_t \to H$ uniformly, with H continuous on the product and each H_t bounded, then*

$$\sup_{c\in\mathcal{C}} \left| \frac{\sup_\lambda\{F(\lambda, c) + tH_t(\lambda, c)\} - V(c)}{t} - \max_{\lambda\in\Gamma(c)} H(\lambda, c) \right| \longrightarrow 0 \quad (t \downarrow 0).$$

The limit $c \mapsto \max_{\lambda\in\Gamma(c)} H(\lambda, c)$ is continuous in c and Lipschitz in H with respect to the uniform norm. The analogous statement holds for minima.

Proof. The function V is continuous because F is uniformly continuous. Fix $\eta > 0$. Multipliers at distance at least η from $\Gamma(c)$ are uniformly separated from the optimum:

$$\inf_{c\in\mathcal{C},\, \lambda\in\Lambda:\mathrm{dist}(\lambda,\Gamma(c))\geq\eta} \{V(c) - F(\lambda, c)\} > 0 \tag{A.12}$$

whenever the set over which the infimum is taken is nonempty. If this failed, there would be a sequence (λ_k, c_k) in that set with $V(c_k) - F(\lambda_k, c_k) \to 0$. By compactness, a subsequence converges to some (λ, c), and by the continuity of F, of V and of $(\lambda, c) \mapsto \mathrm{dist}(\lambda, \Gamma(c))$, the limit satisfies $F(\lambda, c) = V(c)$ and $\mathrm{dist}(\lambda, \Gamma(c)) \geq \eta$, which is impossible.

For every c, choose $\lambda_t(c) \in \Lambda$ at which $F(\cdot, c) + tH_t(\cdot, c)$ is within t^2 of its supremum; since the bounds below are deterministic, this choice need not be measurable in c. Because $\|H_t\|_\infty$ is bounded for small t,

$$0 \leq V(c) - F(\lambda_t(c), c) \leq 2t\|H_t\|_\infty + t^2.$$

Equation (A.12) then gives $\sup_c \mathrm{dist}(\lambda_t(c), \Gamma(c)) \to 0$. Evaluating $F+tH_t$ at a maximizer of H on $\Gamma(c)$ shows that the difference quotient in the lemma is at least $\max_{\Gamma(c)} H - \|H_t - H\|_\infty$, and evaluating it at $\lambda_t(c)$ shows that the quotient is at most $H_t(\lambda_t(c), c) + t$. Because H is uniformly continuous and $\sup_c \mathrm{dist}(\lambda_t(c), \Gamma(c)) \to 0$, this upper bound is at most $\max_{\Gamma(c)} H + o(1)$, uniformly in c, which proves the expansion. The continuity of $c \mapsto \max_{\Gamma(c)} H$ follows from the Hausdorff continuity of Γ and the uniform continuity of H, and the Lipschitz property holds because $|\max_{\Gamma(c)} H - \max_{\Gamma(c)} \widetilde{H}| \leq \|H - \widetilde{H}\|_\infty$ for any two continuous functions H and $\widetilde{H}$ on $\Lambda \times \mathcal{C}$. For minima, apply the result to $-F$, $-H_t$ and $-H$. □

Proof of Theorem 6. Let P_t be an admissible family, so that $\mathcal{S}(P_t) = \mathcal{S}(P_0) + th_t$ with $h_t \to h = (h_p, h_+, h_-)$ in $\mathbb{D}$ and $h \in \mathbb{T}$. By Proposition 1, the multipliers can be restricted to $\Lambda = [0, 1/\varepsilon]$ for every P_t. By the Cauchy–Schwarz inequality, the $L_2(P_0)$ continuity assumed in the theorem implies that $P_0 f^{\pm}_{\lambda,c}$ is continuous on $\mathcal{K}$.

Step 1: expansion of the worst-case rates. Write

$$\Psi_0^+(\lambda, c) = \frac{P_0 f^+_{\lambda,c} - \lambda\varepsilon}{p_0}, \qquad \Psi_0^-(\lambda, c) = \frac{P_0 f^-_{\lambda,c} + \lambda\varepsilon}{1 - p_0},$$

and let $\Psi_t^{\pm}$ denote the same expressions with P_0 and p_0 replaced by P_t and P_tY. By Proposition 1, $B_t(c) := \underline{\beta}_\varepsilon(c; P_t)$ is the maximum of $\Psi_t^+(\cdot, c)$ over Λ, and $A_t(c) := \overline{\alpha}_\varepsilon(c; P_t)$ is the minimum of $\Psi_t^-(\cdot, c)$. Writing each difference $\Psi_t^{\pm} - \Psi_0^{\pm}$ over a common denominator, and using the fact that P_tY is bounded away from zero and one for small t and that h_t is bounded, we obtain, uniformly on $\mathcal{K}$,

$$\frac{\Psi_t^+ - \Psi_0^+}{t} \longrightarrow D_h^+(\lambda, c) = \frac{h_+(\lambda, c) - \Psi_0^+(\lambda, c)h_p}{p_0},$$
$$\frac{\Psi_t^- - \Psi_0^-}{t} \longrightarrow D_h^-(\lambda, c) = \frac{h_-(\lambda, c) + \Psi_0^-(\lambda, c)h_p}{1 - p_0}.$$

Both limits are continuous on $\mathcal{K}$. Applying Lemma 4 to the maximum and to the minimum therefore gives

$$\|B_t - B_0 - t\dot{B}_h\|_{\mathcal{C}} = o(t), \qquad \|A_t - A_0 - t\dot{A}_h\|_{\mathcal{C}} = o(t). \tag{A.13}$$

Since Ψ_0^+ equals $B_0(c)$ on $\Gamma_+(c)$ and Ψ_0^- equals $A_0(c)$ on $\Gamma_-(c)$, the derivatives are those in (A.7) and (A.8), and by Lemma 4 they are continuous in c.

Step 2: the inverse of the worst-case FPR. Let $c_{\alpha,t} = \min\{c \in \mathcal{C} : A_t(c) \le \alpha\}$. Each A_t is nonincreasing and right-continuous, as in Lemma 2. By (A.13), $\|A_t - A_0\|_{\mathcal{C}} = O(t)$, so the endpoint inequalities in Assumption 1(ii) and the argument that leads to (A.4) give

$$\sup_\alpha |c_{\alpha,t} - c_\alpha| = O(t),$$

and all these thresholds lie in the interior of $\mathcal{C}$ for small t. Let $R_t = A_t - A_0 - t\dot{A}_h$, so that $\|R_t\|_\infty = o(t)$ by (A.13). Since $A_0 + t\dot{A}_h$ is continuous, every jump of A_t in the interior of $\mathcal{C}$ is at most $2\|R_t\|_\infty = o(t)$. By the minimality of $c_{\alpha,t}$ and right-continuity,

$$A_t(c_{\alpha,t}) \le \alpha \le A_t(c_{\alpha,t}-).$$

Consequently,

$$\sup_\alpha |A_t(c_{\alpha,t}) - \alpha| = o(t). \tag{A.14}$$

By (A.13) and the uniform continuity of $\dot{A}_h$,

$$\sup_\alpha \left| A_t(c_{\alpha,t}) - A_0(c_{\alpha,t}) - t\dot{A}_h(c_\alpha) \right| = o(t).$$

Since $A_0(c_\alpha) = \alpha$, this bound, the bound on $c_{\alpha,t} - c_\alpha$ and the uniform continuity of A_0' give

$$A_t(c_{\alpha,t}) - \alpha = A_0'(c_\alpha)(c_{\alpha,t} - c_\alpha) + t\dot{A}_h(c_\alpha) + o(t)$$

uniformly in α. Dividing by t, using (A.14), and dividing by $A_0'(c_\alpha)$, whose absolute value is at least $\underline{L}_\alpha$, proves the formula for $\dot{c}_h$ in (A.9), uniformly in α.

Step 3: the worst-case ROC curve. Because P_t satisfies the endpoint inequalities, $c_{\alpha,t}$ is also the smallest threshold in $[0, 1]$ that is feasible at level α under P_t, so Lemma 2 gives $g_\varepsilon(\alpha; P_t) = B_t(c_{\alpha,t})$. Hence (A.13), the bound on $c_{\alpha,t} - c_\alpha$, and the continuity of B_0' and $\dot{B}_h$ give

$$g_\varepsilon(\alpha; P_t) - g_\varepsilon(\alpha; P_0) = t\dot{B}_h(c_\alpha) + B_0'(c_\alpha)(c_{\alpha,t} - c_\alpha) + o(t)$$

uniformly in α, and substituting $\dot{c}_h$ gives the expansion of the worst-case ROC curve stated in the theorem. The derivative map $h \mapsto \dot{g}_h$ is continuous, in fact Lipschitz, because the maximum and the minimum over a fixed compact set are Lipschitz and the ratios $1/A_0'$ and B_0'/A_0' are bounded on $\mathcal{C}$. Since these expansions hold for every admissible family P_t, the map $\mathcal{S}(P) \mapsto g_\varepsilon(\cdot; P)$ is Hadamard directionally differentiable at $\mathcal{S}(P_0)$ tangentially to $\mathbb{T}$, in the norm of $\mathbb{D}$. If the sets $\Gamma_\pm(c)$ are singletons, the derivatives in (A.7) to (A.9) are linear in h. □

Proof of (A.11). The classes $\{f^+_{\lambda,c}\}$ and $\{f^-_{\lambda,c}\}$ differ from $\mathcal{F}^\pm$ only by the constant factors p_0 and $1-p_0$, so they have bounded envelopes and satisfy a polynomial covering bound as in Assumption 1(i); the same holds after the function Y is added. The finite-dimensional distributions of the centered and scaled empirical averages converge by the multivariate central limit theorem. For tightness, consider differences $f-g$ of two functions in the same class with $\|f-g\|_{L_2(P_0)} \le \delta$. Approximating each function pointwise by members of the countable subclass given by pointwise measurability, and applying bounded convergence, shows that the supremum of the empirical process over these differences is bounded by its supremum over countably many differences with $L_2(P_0)$ norm less than 2δ. This countable class satisfies a polynomial covering bound and has envelope at most two, and both it and the class of its squares are measurable. After the envelope is rescaled, the local maximal inequality of van der Vaart and Wellner (2011, Theorem 2.1) bounds the limit superior, as $n \to \infty$, of the expected supremum by a constant times

$$\int_0^{2\delta} \sqrt{1 + v \log(C/u)}\, du,$$

where $C \ge 1$ is a fixed covering constant. This integral tends to zero as $\delta \downarrow 0$, and the remaining term in the inequality tends to zero as $n \to \infty$ for each fixed δ. Suprema over the countable classes are measurable, and outer probabilities can be used for the original processes. Because $\mathcal{K}$ is compact and $(\lambda, c) \mapsto f^\pm_{\lambda,c}$ is continuous into $L_2(P_0)$, the index set is totally bounded, so this asymptotic equicontinuity implies tightness in $\mathbb{D}$ (van der Vaart and Wellner, 1996, Theorems 1.5.4 and 1.5.7), and the limiting Gaussian process has a version with continuous paths on $\mathcal{K}$. Thus

$$\sqrt{n}\{\mathcal{S}(\widehat{\mathbb{P}}_n) - \mathcal{S}(P_0)\} \rightsquigarrow \mathbb{G}, \qquad \mathbb{P}(\mathbb{G} \in \mathbb{T}) = 1.$$

Samples in which a class is absent or $\widehat{\mathbb{P}}_n$ violates one of the endpoint inequalities have probability tending to zero, because $0 < p_0 < 1$ and $\mathbb{P}(\delta_{\alpha,n} < b_C) \to 1$ (Step 3 of the proof of Theorem 1); on them the estimated worst-case ROC curve may be set to any fixed bounded function. For every other sample, $\mathcal{S}(\widehat{\mathbb{P}}_n)$ belongs to the domain of the map $\mathcal{S}(P) \mapsto g_\varepsilon(\cdot; P)$. The conditions of Fang and Santos (2019, Theorem 2.1) therefore hold: $\mathbb{D}$ and $\ell^\infty([\alpha_l, \alpha_u])$ are Banach spaces, the map is Hadamard directionally differentiable at $\mathcal{S}(P_0)$ tangentially to $\mathbb{T}$ by Theorem 6, and the limit $\mathbb{G}$ is tight and concentrated on $\mathbb{T}$. This proves (A.11). □

Lemma 5 (Verification for the logistic score)**.** *Under the conditions of Remark 3, the two continuity conditions of Theorem 6 hold, and on $\mathcal{C}$ both worst-case rates are continuously differentiable with strictly negative derivatives bounded away from zero.*

Proof. Let H_y be the CDF of S given $Y = y$, as in Remark 1, and $h_y = H_y'$ its density; write $a = a(c)$, and put $r_1 = \varepsilon/p_0$ and $r_0 = \varepsilon/(1-p_0)$. Because the densities are strictly

positive and (A.10) holds, there are unique finite $\omega_1(a), \omega_0(a) > 0$ that solve

$$\int_0^{\omega_1(a)} u h_1(a+u)\,du = r_1, \qquad \int_0^{\omega_0(a)} u h_0(a-u)\,du = r_0. \tag{A.15}$$

Indeed, each integral is continuous and strictly increasing in its upper limit, equals zero when the upper limit is zero, and tends to the corresponding expectation in (A.10) as the upper limit tends to infinity. Since the integrals are jointly continuous in a and the upper limit, and each root is bracketed by two strict inequalities, ω_y is continuous on the compact set of values of a.

For $\lambda > 0$, the derivative with respect to λ of the dual objective for the TPR, conditional on $Y = 1$, is

$$\mathbb{E}[(S-a)\mathbf{1}\{a < S < a + 1/\lambda\} \mid Y = 1] - r_1.$$

Here we use the fact that S has no atom at the endpoint $a + 1/\lambda$. The derivative is strictly decreasing in λ, from a positive value to $-r_1$, so its zero is the unique maximizer, $\lambda_+(c) = 1/\omega_1(a(c))$. The derivative of the dual objective for the FPR is

$$r_0 - \mathbb{E}[(a-S)\mathbf{1}\{a - 1/\lambda < S < a\} \mid Y = 0],$$

which is strictly increasing and passes through zero; the unique minimizer is therefore $\lambda_-(c) = 1/\omega_0(a(c))$. By Proposition 1, both lie in Λ, and since they are continuous in c, the Hausdorff continuity condition holds. The functions $f^{\pm}_{\lambda,c}$ are bounded and continuous in (λ, c) at every (x, y), so by dominated convergence they are continuous in $L_2(P_0)$, including at $\lambda = 0$.

Set $z_1(a) = a + \omega_1(a)$ and $z_0(a) = a - \omega_0(a)$. Substituting (A.15) into the dual objectives gives

$$B_0(c) = 1 - H_1(z_1(a(c))), \qquad A_0(c) = 1 - H_0(z_0(a(c))).$$

Thus, as described in Section 2.2, the worst case for the TPR moves the sick cases with $a < S \le z_1(a)$, the closest ones above the threshold, onto it. The worst case for the FPR moves the healthy cases with $z_0(a) < S \le a$, the closest ones below the threshold, just above it. By (A.15), each of these moves uses up the budget. To differentiate B_0 and A_0 without assuming that the densities are differentiable, write the first equation in (A.15) as $\int_a^{z_1} (v-a)h_1(v)\,dv = r_1$ and the second as $\int_{z_0}^{a} (a-v)h_0(v)\,dv = r_0$. These integrals are continuously differentiable in their endpoints, with nonzero derivatives with respect to z_1 and z_0, and implicit differentiation gives

$$\frac{dB_0}{da} = -\frac{H_1(a+\omega_1(a)) - H_1(a)}{\omega_1(a)},$$
$$\frac{dA_0}{da} = -\frac{H_0(a) - H_0(a-\omega_0(a))}{\omega_0(a)}.$$

Each ratio is the average of a continuous, strictly positive density over an interval whose endpoints are continuous in a; it is therefore continuous and strictly positive, with a positive minimum and a finite maximum on the compact set of values of a. Finally, $a'(c) = 1/(\|w\|_2 c(1-c))$ is positive and bounded above and away from zero on $\mathcal{C} \subset (0,1)$, which proves the claim about the derivatives with respect to c. □

Replacement decisions and selected thresholds.

Lemma 6 (Dirichlet posterior calculation and concentration)**.** *For a fixed doctor, suppose that the n_j pairs $(Y_i, \widehat{Y}_{i,j})$ in the doctor's history are independent and identically distributed, with $p_j = \mathbb{P}(Y_i = 1) \in [p_*, 1-p_*]$ for some $p_* \in (0, 1/2]$, and that the four cell probabilities have a fixed symmetric Dirichlet prior with parameter $a_0 > 0$. Then the doctor's FPR and TPR are independent under the posterior and have the distributions in (2.17), and (2.18) is exactly the posterior probability of the event $\{\alpha_H \geq \widehat{\overline{\alpha}}_\varepsilon(c),\ \beta_H \leq \widehat{\underline{\beta}}_\varepsilon(c)\}$. Moreover, Assumption 3 holds with, for example,*

$$C_j = C(a_0) := \max\{8, \exp(a_0/4)\} + 2, \qquad \zeta_j = \zeta(p_*) := \min\{p_*/16, p_*^2/2\}. \tag{A.16}$$

Proof. The multinomial likelihood and the Dirichlet prior give a Dirichlet posterior for the four cell probabilities. We represent this posterior by four independent Gamma variables, each divided by the sum of the four, where the shape parameters of the Gamma variables are the four posterior cell parameters. The sum cancels in the ratios that define the FPR and the TPR, and the two ratios involve disjoint pairs of the Gamma variables. Hence the FPR and the TPR are independent under the posterior and have the Beta distributions in (2.17). By independence, the posterior probability of the event in the lemma is the product in (2.18).

We next bound the tails of the posterior, conditional on the observed counts. Let $Z \sim \mathrm{Beta}(a, b)$, write $s = a + b$ and $\mu = a/s$, and represent Z as $G_1/(G_1 + G_2)$ with independent Gamma variables of unit scale. For $\mu < x < 1$, the Chernoff bound gives

$$\mathbb{P}(Z \geq x) \leq \inf_{0<v<1/(1-x)} \{1 - v(1-x)\}^{-a}(1 + vx)^{-b}.$$

Taking $v = (x - \mu)/\{x(1-x)\}$ yields $\mathbb{P}(Z \geq x) \leq \exp\{-s\,\mathrm{kl}(\mu, x)\}$, where

$$\mathrm{kl}(\mu, x) = \mu \log(\mu/x) + (1-\mu)\log\{(1-\mu)/(1-x)\}.$$

For $0 < x < \mu$, the same calculation with $v < 0$ bounds $\mathbb{P}(Z \leq x)$. Since $\partial_x \mathrm{kl}(\mu, x) = (x - \mu)/\{x(1-x)\}$ and $x(1-x) \leq 1/4$, integrating from μ to x gives $\mathrm{kl}(\mu, x) \geq 2(x-\mu)^2$ for x on either side of μ. For $u > 0$ with $\mu + u \geq 1$ or $\mu - u \leq 0$, the corresponding tail probability is zero, so in all cases

$$\mathbb{P}(|Z - \mu| > u) \leq 2\exp(-2su^2), \qquad u > 0. \tag{A.17}$$

Since (A.17) is centered at the mean of the Beta distribution, it remains to relate the posterior means to the true error rates.

Let $N_{j0} = \mathrm{FP}_j + \mathrm{TN}_j$, $N_{j1} = \mathrm{TP}_j + \mathrm{FN}_j$ and $m_j = \min(N_{j0}, N_{j1})$. Conditional on these class counts, FP_j and TP_j are binomial under the sampling assumption of the lemma. If $m_j > 0$, let $\widetilde{\theta}_j = (\mathrm{FP}_j/N_{j0}, \mathrm{TP}_j/N_{j1})$, and let μ_j be the vector of posterior means. Then

$$\|\mu_j - \widetilde{\theta}_j\|_\infty \leq b_j := \frac{a_0}{m_j + 2a_0}.$$

The inequality of Hoeffding (1963) for bounded variables gives

$$\mathbb{P}\big(\|\widetilde{\theta}_j - \theta_{H,j}\|_\infty > u \mid N_{j0}, N_{j1}\big) \leq 4\exp(-2m_j u^2). \tag{A.18}$$

If $t \geq 4b_j$, then on the event $\{\|\widetilde{\theta}_j - \theta_{H,j}\|_\infty \leq t/4\}$ the posterior mean is within $t/2$ of $\theta_{H,j}$. Applying (A.17) to the two posterior coordinates then gives

$$T_j(t) \leq 4\exp\{-(m_j + 2a_0)t^2/2\}$$

on this event. By (A.18), the conditional probability of its complement is at most $4\exp(-m_j t^2/8)$. Therefore

$$\mathbb{E}[T_j(t) \mid N_{j0}, N_{j1}] \leq 8\exp(-m_j t^2/8).$$

If $t < 4b_j$, then $T_j(t) \leq 1$ and

$$\frac{m_j t^2}{8} < \frac{2m_j a_0^2}{(m_j + 2a_0)^2} \leq \frac{a_0}{4},$$

so $T_j(t) \leq e^{a_0/4}\exp(-m_j t^2/8)$. If one class is absent, then $m_j = 0$, and the bound $T_j(t) \leq 1$ still applies. Hence, for all class counts and $0 < t \leq 1$,

$$\mathbb{E}[T_j(t) \mid N_{j0}, N_{j1}] \leq C_a \exp(-m_j t^2/8), \qquad C_a = \max\{8, e^{a_0/4}\}. \tag{A.19}$$

Finally, N_{j1} is binomial with success probability p_j, and Hoeffding's inequality applied to the class counts yields

$$\mathbb{P}(m_j < n_j p_*/2) \leq 2\exp(-n_j p_*^2/2).$$

Using (A.19) on the complement of this event and $T_j(t) \leq 1$ on the event itself, we obtain

$$\begin{aligned}\mathbb{E}[T_j(t)] &\leq C_a \exp(-n_j p_* t^2/16) + 2\exp(-n_j p_*^2/2) \\ &\leq (C_a + 2)\exp\{-n_j t^2 \min(p_*/16, p_*^2/2)\},\end{aligned}$$

where the last step uses $t \leq 1$. This is the bound in Assumption 3, with the constants in (A.16). □

Proof of Theorem 2. We work on the event $\mathcal{E}_n = \{r_n \leq \delta_0/4\}$. For a candidate c, we call $\{\theta_H : \alpha_H \geq \widehat{\overline{\alpha}}_\varepsilon(c),\ \beta_H \leq \widehat{\underline{\beta}}_\varepsilon(c)\}$ the rectangle of c. It consists of the error rates of doctors whom c dominates under every distribution in $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_n)$, and its posterior probability is $\widehat{q}_j^{\varepsilon}(c)$. The neighborhood

$$K_j = \{\theta_H : \|\theta_H - \theta_{H,j}\|_\infty \leq t_0\}$$

of the doctor's true error rates has posterior probability $1 - T_j(t_0)$.

Consider first a dominated doctor, with $\beta_{H,j} \leq g_\varepsilon(\alpha_{H,j}; P_0) - \delta_0$ as in part (i), and let $\alpha^* = \alpha_{H,j} - t_0$, which lies in $[\alpha_l, \alpha_u]$ because $t_0 \leq d_H/2$. By Lemma 2, the estimated worst-case ROC curve computed over $\mathcal{C}_n$ is attained, so there is a candidate c^* with

$$\widehat{\overline{\alpha}}_\varepsilon(c^*) \leq \alpha^*, \qquad \widehat{\underline{\beta}}_\varepsilon(c^*) = \widehat{g}_\varepsilon(\alpha^*).$$

Every $\theta_H \in K_j$ satisfies $\alpha_H \geq \alpha^* \geq \widehat{\overline{\alpha}}_\varepsilon(c^*)$. In addition, by the definition of r_n and the

Lipschitz condition in Assumption 2(ii),

$$\begin{aligned}\widehat{g}_\varepsilon(\alpha^*) &\geq g_\varepsilon(\alpha_{H,j}; P_0) - L_g t_0 - r_n \\ &\geq \beta_{H,j} + \delta_0 - L_g t_0 - r_n > \beta_{H,j} + t_0 \geq \beta_H,\end{aligned}$$

because $(1 + L_g)t_0 + r_n \leq \delta_0/2$ on $\mathcal{E}_n$. Thus the rectangle of c^* contains K_j, and since $\widehat{q}^{\varepsilon}_{\max,j}$ is the maximum over all candidates, $\widehat{q}^{\varepsilon}_{\max,j} \geq 1 - T_j(t_0)$.

Consider next a doctor who is not dominated, with $\beta_{H,j} \geq g_\varepsilon(\alpha_{H,j}; P_0) + \delta_0$ as in part (ii), and suppose that the rectangle of some candidate c intersects K_j. Then $\widehat{\overline{\alpha}}_\varepsilon(c) \leq \alpha_{H,j} + t_0$ and $\widehat{\underline{\beta}}_\varepsilon(c) \geq \beta_{H,j} - t_0$. But c is then feasible at level $\alpha_{H,j} + t_0$ under $\widehat{\mathbb{P}}_n$, a level that also lies in $[\alpha_l, \alpha_u]$, so the definition of r_n and Assumption 2(ii) give

$$\begin{aligned}\widehat{\underline{\beta}}_\varepsilon(c) &\leq \widehat{g}_\varepsilon(\alpha_{H,j} + t_0) \\ &\leq g_\varepsilon(\alpha_{H,j}; P_0) + L_g t_0 + r_n < \beta_{H,j} - t_0,\end{aligned}$$

a contradiction. Hence the rectangle of every candidate is disjoint from K_j and has posterior probability at most $T_j(t_0)$, so that $\widehat{q}^{\varepsilon}_{\max,j} \leq T_j(t_0)$.

On $\mathcal{E}_n$, the two bounds show that a dominated doctor is retained only if $T_j(t_0) > 1 - \tau$ and that a doctor who is not dominated is replaced only if $T_j(t_0) \geq \tau$. Since every doctor is of one of these two kinds by Assumption 2(i), the event that at least one decision is wrong is contained, on $\mathcal{E}_n$, in $\bigcup_j \{T_j(t_0) \geq \bar{\tau}\}$. By Markov's inequality,

$$\mathbb{P}(\text{at least one wrong decision}) \leq \mathbb{P}(\mathcal{E}_n^c) + \frac{1}{\bar{\tau}} \sum_{j=1}^{J} \mathbb{E}[T_j(t_0)],$$

which, by Assumption 3, proves (3.7). The same tail bound gives $T_j(t_0) \to_p 0$, and with it the limits of $\widehat{q}^{\varepsilon}_{\max,j}$ stated in the theorem. □

Proof of Theorem 3. For every candidate c and every $Q \in \mathcal{B}_\varepsilon(P_0)$, the definitions of the worst-case FPR and of $e_{\alpha,n}$ give $\mathrm{FPR}_Q(c) \leq \overline{\alpha}_\varepsilon(c; P_0) \leq \widehat{\overline{\alpha}}_\varepsilon(c) + e_{\alpha,n}$, and in the same way $\mathrm{TPR}_Q(c) \geq \underline{\beta}_\varepsilon(c; P_0) \geq \widehat{\underline{\beta}}_\varepsilon(c) - e_{\beta,n}$. With $c = \widehat{c}_j$, this is (3.10). Every pair $\theta_H = (\alpha_H, \beta_H)$ in the rectangle of $\widehat{c}_j$, defined in the proof of Theorem 2, satisfies

$$\widehat{\overline{\alpha}}_\varepsilon(\widehat{c}_j) \leq \alpha_H, \qquad \widehat{\underline{\beta}}_\varepsilon(\widehat{c}_j) \geq \beta_H.$$

By (3.10), the two inequalities in the event displayed in the theorem then hold for this pair, simultaneously for all $Q \in \mathcal{B}_\varepsilon(P_0)$. The rectangle of $\widehat{c}_j$, whose posterior probability is at least τ, is therefore contained in that event.

If $T_j(t) < \tau$, this rectangle cannot be disjoint from the t-neighborhood $\{\theta_H : \|\theta_H - \theta_{H,j}\|_\infty \leq t\}$ of the doctor's true error rates, because the posterior probability of their intersection is at least $\tau - T_j(t) > 0$. Taking a pair θ_H in the intersection, we obtain

$$\widehat{\overline{\alpha}}_\varepsilon(\widehat{c}_j) \leq \alpha_{H,j} + t, \qquad \widehat{\underline{\beta}}_\varepsilon(\widehat{c}_j) \geq \beta_{H,j} - t.$$

Substituting these inequalities into (3.10) proves (3.11).

If the square root in (3.9) is less than one, Markov's inequality, Assumption 3 and a

union bound give

$$\mathbb{P}\{\max_j T_j(t_n(\gamma)) > \tau/2\} \le \frac{2JC_0}{\tau}\exp\{-\zeta_0 n_{\min} t_n(\gamma)^2\} \le \gamma.$$

If the square root is at least one, then $t_n(\gamma) = 1$ and $T_j(1) = 0$, because the true error rates and every pair in the support of the posterior lie in $[0,1]^2$. On the complement of the event in the last display, $T_j(t_n(\gamma)) < \tau$ for every j, so the rectangle of every selected threshold intersects the $t_n(\gamma)$-neighborhood of the doctor's true error rates, and the last statement of the theorem follows. □

Welfare comparisons.

Proof of Corollary 1. Consider the event on which the last statement of Theorem 3 holds; its sampling probability is at least $1-\gamma$. On this event, (3.11) with $t = t_n(\gamma)$ gives, for every replaced doctor and every $Q \in \mathcal{B}_\varepsilon(P_0)$,

$$\begin{aligned} W_M(\widehat{c}_j; Q) - W_{H,j} &= \phi[\mathrm{TPR}_Q(\widehat{c}_j) - \beta_{H,j}] - \eta[\mathrm{FPR}_Q(\widehat{c}_j) - \alpha_{H,j}] \\ &\ge -\phi e_{\beta,n} - \eta e_{\alpha,n} - (\phi+\eta)t_n(\gamma). \end{aligned}$$

Retained doctors contribute nothing to the difference from retaining all doctors, so summing over the replaced doctors proves (3.13) on this event. □

Proof of Theorem 4. Part (i). Let Q be as in the theorem. A threshold whose worst-case FPR is at most α also has FPR at most α under Q, and its TPR under Q is at least its worst-case TPR. Thus $g_\varepsilon(\cdot; P_0) \le g_Q$ and $\mathcal{R}_\varepsilon(P_0) \subseteq \mathcal{R}^*(Q)$. For a doctor in both sets,

$$\begin{aligned} W_M(c_j^*; Q) - W_M(c_{\varepsilon,j}; Q) &= \phi\{g_Q(\alpha_{H,j}) - \mathrm{TPR}_Q(c_{\varepsilon,j})\} - \eta\{\alpha_{H,j} - \mathrm{FPR}_Q(c_{\varepsilon,j})\} \\ &\le \phi\{g_Q(\alpha_{H,j}) - g_\varepsilon(\alpha_{H,j}; P_0)\} \le \phi\Delta_\varepsilon(Q). \end{aligned}$$

The first inequality holds because, by Definition 4, $\mathrm{TPR}_Q(c_{\varepsilon,j}) \ge \underline{\beta}_\varepsilon(c_{\varepsilon,j}; P_0) = g_\varepsilon(\alpha_{H,j}; P_0)$ and $\mathrm{FPR}_Q(c_{\varepsilon,j}) \le \overline{\alpha}_\varepsilon(c_{\varepsilon,j}; P_0) = \alpha_{H,j}$, so the FPR term is nonpositive. The second holds because $\alpha_{H,j} \in [\alpha_l, \alpha_u]$. If $j \in \mathcal{R}^*(Q) \setminus \mathcal{R}_\varepsilon(P_0)$, only the benchmark replaces the doctor, and its threshold has the doctor's FPR, so the welfare difference for this doctor is exactly $\phi\{g_Q(\alpha_{H,j}) - \beta_{H,j}\}$. Since $\beta_{H,j} \ge g_\varepsilon(\alpha_{H,j}; P_0)$ for such a doctor,

$$0 < g_Q(\alpha_{H,j}) - \beta_{H,j} \le \Delta_\varepsilon(Q).$$

By (3.16) evaluated at $\Delta_\varepsilon(Q)$, the share of such doctors is at most $C_M\Delta_\varepsilon(Q)^{\gamma_M}$, and adding the two contributions gives (3.19). If $\Delta_\varepsilon(Q) = 0$, the last display shows that no doctor is replaced only by the benchmark, so the two replacement sets coincide, and the difference for every doctor in them is nonpositive; (3.19) therefore holds in this case as well.

Part (ii), Step 1: the posterior event. With $T_j(t)$ as in (3.5) and t as in (3.18), define

$$\mathcal{E}(t) := \{\max_{j \le J} T_j(t) \le \bar{\tau}/2\}.$$

By Markov's inequality, Assumption 3 and a union bound,

$$\mathbb{P}\{\mathcal{E}(t)^c\} \le \frac{2JC_0}{\bar{\tau}} \exp(-\zeta_0 n_{\min} t^2) = \gamma.$$

The event $\mathcal{E}(t)$ does not depend on Q. On $\mathcal{E}(t)$, every median $a_{n,j}$ lies in $[\alpha_{H,j} - t, \alpha_{H,j} + t]$; otherwise at least half of the posterior mass of α_H would lie outside this interval, whereas $T_j(t) \le \bar{\tau}/2 \le 1/4$. Because $t \le d_H$, the medians and the points $\alpha_{H,j} \pm t$ all lie in $[\alpha_l, \alpha_u]$.

Step 2: locating extra and missed replacements. We work on $\mathcal{E}(t)$. Suppose first that $g_\varepsilon(\alpha_{H,j}; P_0) - \beta_{H,j} > h_n$, and let $\alpha_* := \alpha_{H,j} - t$. Then, by the definition of r_n and Assumption 2(ii),

$$\widehat{g}_\varepsilon(\alpha_*) \ge g_\varepsilon(\alpha_{H,j}; P_0) - L_g t - r_n > \beta_{H,j} + t.$$

By Lemma 2, some candidate attains $\widehat{g}_\varepsilon(\alpha_*)$; its estimated worst-case FPR is at most $\alpha_{H,j} - t$, and its estimated worst-case TPR exceeds $\beta_{H,j} + t$. The rectangle of this candidate, defined in the proof of Theorem 2, therefore contains the t-neighborhood of $\theta_{H,j}$, so its posterior probability is at least $1 - T_j(t) > \tau$, and doctor j is replaced.

Conversely, if doctor j is replaced, the rectangle of $\widehat{c}_j$ has posterior probability at least $\tau > T_j(t)$. Hence it intersects the t-neighborhood of $\theta_{H,j}$, so that

$$\widehat{\overline{\alpha}}_\varepsilon(\widehat{c}_j) \le \alpha_{H,j} + t, \qquad \widehat{\underline{\beta}}_\varepsilon(\widehat{c}_j) \ge \beta_{H,j} - t. \tag{A.20}$$

Then $\widehat{c}_j$ is feasible at level $\alpha_{H,j} + t$ under $\widehat{\mathbb{P}}_n$, so $\widehat{\underline{\beta}}_\varepsilon(\widehat{c}_j) \le \widehat{g}_\varepsilon(\alpha_{H,j} + t)$, and it follows that

$$\beta_{H,j} \le g_\varepsilon(\alpha_{H,j}; P_0) + r_n + (1 + L_g)t = g_\varepsilon(\alpha_{H,j}; P_0) + h_n.$$

For $j \in \mathcal{A}_n$, the benchmark does not replace the doctor, so $\beta_{H,j} \ge g_Q(\alpha_{H,j}) \ge g_\varepsilon(\alpha_{H,j}; P_0)$, and the last display gives

$$0 \le \beta_{H,j} - g_Q(\alpha_{H,j}) \le h_n.$$

For $j \in \mathcal{M}_n$, the implemented rule does not replace the doctor, so the first implication gives $g_\varepsilon(\alpha_{H,j}; P_0) - \beta_{H,j} \le h_n$, and hence

$$0 < g_Q(\alpha_{H,j}) - \beta_{H,j} \le \Delta_\varepsilon(Q) + h_n.$$

Evaluating (3.16) at h_n and at $\Delta_\varepsilon(Q) + h_n$ now gives (3.21); the separation δ_0 in Assumption 2(i) is not needed.

Step 3: threshold differences among common replacements. Write $A_{n,j} = \widehat{\overline{\alpha}}_\varepsilon(\widehat{c}_j)$ and $B_{n,j} = \widehat{\underline{\beta}}_\varepsilon(\widehat{c}_j)$. For each $j \in \mathcal{I}_n$, (3.10) and the definition of $c_j^*(Q)$ imply

$$W_M(c_j^*; Q) - W_M(\widehat{c}_j; Q) \le \phi\{g_Q(\alpha_{H,j}) - B_{n,j}\} + \eta\{A_{n,j} - \alpha_{H,j}\} + \phi e_{\beta,n} + \eta e_{\alpha,n}.$$

Because $|a_{n,j} - \alpha_{H,j}| \le t$ on $\mathcal{E}(t)$, the definitions of $\Delta_\varepsilon(Q)$ and r_n and Assumption 2(ii) give

$$g_Q(\alpha_{H,j}) \le \Delta_\varepsilon(Q) + g_\varepsilon(\alpha_{H,j}; P_0) \le \Delta_\varepsilon(Q) + \widehat{g}_\varepsilon(a_{n,j}) + r_n + L_g t.$$

Moreover, $\eta(a_{n,j} - \alpha_{H,j}) \le \eta t$. Substituting these bounds gives

$$\begin{aligned} W_M(c_j^*; Q) - W_M(\widehat{c}_j; Q) \le{}& \phi\Delta_\varepsilon(Q) + v_n \\ &+ \phi\{\widehat{g}_\varepsilon(a_{n,j}) - B_{n,j}\} + \eta\{A_{n,j} - a_{n,j}\}. \end{aligned}$$

We bound the sum of the last two terms by its positive part and sum over $j \in \mathcal{I}_n$. Since $\mathcal{I}_n \subseteq \mathcal{R}_\varepsilon^{(n)}$, the contribution of the common replacements to (3.15) is at most

$$\widehat{\pi}_n\{\phi\Delta_\varepsilon(Q) + v_n\} + \widehat{K}_n.$$

Step 4: the remaining contributions. For $j \in \mathcal{M}_n$, the definition of $c_j^*(Q)$ and Step 2 give

$$W_M(c_j^*; Q) - W_{H,j} = \phi\{g_Q(\alpha_{H,j}) - \beta_{H,j}\} \leq \phi\{\Delta_\varepsilon(Q) + h_n\}.$$

For $j \in \mathcal{A}_n$, (A.20) and (3.10) give

$$W_{H,j} - W_M(\widehat{c}_j; Q) \leq \phi(e_{\beta,n} + t) + \eta(e_{\alpha,n} + t) = u_n.$$

Multiplying these bounds by the bounds on $|\mathcal{M}_n|/J$ and $|\mathcal{A}_n|/J$ in (3.21) and adding the contribution of the common replacements, we obtain (3.20) from (3.15). All the inequalities hold on the single event $\mathcal{E}(t)$, for every $Q \in \mathcal{B}_\varepsilon(P_0)$ at which the benchmark is defined and Assumption 4 holds, with the constants associated with that Q. □

Write $\bar{r}_n := \sqrt{\log n/n} + n_{\min}^{-1/2}$.

Proposition 4 (Estimation accuracy and the welfare remainder)**.** *Suppose that the conditions of Theorem 4(ii) other than $t \leq \min\{1, d_H\}$ hold with J and $\varepsilon > 0$ fixed and with a fixed Q at which the benchmark is defined and Assumption 4 holds, and that $n_{\min} \to \infty$ and*

$$r_n + e_{\alpha,n} + e_{\beta,n} = O_p\left(\sqrt{\frac{\log n}{n}}\right).$$

Then the positive part of the remainder satisfies

$$\left[\mathrm{Regret}(\varepsilon, Q) - \widehat{K}_n - \phi\widehat{\pi}_n\Delta_\varepsilon(Q)\right]_+ = O_p\left(\bar{r}_n + \{\Delta_\varepsilon(Q) + \bar{r}_n\}^{1+\gamma_M}\right). \tag{A.21}$$

Moreover, in (3.20) the term due to missed replacements is $O_p(\{\Delta_\varepsilon(Q) + \bar{r}_n\}^{1+\gamma_M})$ and the term due to extra replacements is $O_p(\bar{r}_n^{1+\gamma_M})$.

For fixed ε and Q with $\Delta_\varepsilon(Q) > 0$, the right side of (A.21) does not tend to zero, so the bound on the remainder is then only $O_p(1)$. The estimation errors r_n, $e_{\alpha,n}$ and $e_{\beta,n}$ depend on the fitted score, on the complexity of the classes of dual functions and on how the distances are computed, and through these possibly on the dimension d of X.

Proof of Proposition 4. Fix $\gamma \in (0, 1)$. Because J, C_0 and ζ_0 are fixed, the quantity t in (3.18) is $O(n_{\min}^{-1/2})$, so the remaining condition $t \leq \min\{1, d_H\}$ of Theorem 4(ii) holds for all sufficiently large $n_{\min}$. The assumed rates then imply $h_n, u_n, v_n = O_p(\bar{r}_n)$. Subtracting $\widehat{K}_n + \phi\widehat{\pi}_n\Delta_\varepsilon(Q)$ from (3.20), taking the positive part and using $0 \leq \widehat{\pi}_n \leq 1$, we find that, on the event of probability at least $1 - \gamma$ on which (3.20) holds, the left side of (A.21) is at most

$$v_n + \phi C_M\{\Delta_\varepsilon(Q) + h_n\}^{1+\gamma_M} + C_M u_n h_n^{\gamma_M}.$$

The first term is $O_p(\bar{r}_n)$, the second is $O_p(\{\Delta_\varepsilon(Q) + \bar{r}_n\}^{1+\gamma_M})$, and the third is $O_p(\bar{r}_n^{1+\gamma_M})$. Since γ is arbitrary, (A.21) follows from these bounds and the tightness of $(n/\log n)^{1/2}(r_n + e_{\alpha,n} + e_{\beta,n})$. The second and third terms are the terms due to missed and extra replacements in (3.20), which proves the last statement. □

Calculations for Example 2.

Proof of the claims in Example 2. The worst-case ROC curve. As in Example 2, for thresholds with $\log\{c/(1-c)\} \in [0,1]$, write $u = 1 - \log\{c/(1-c)\}$; at these thresholds the nominal TPR and FPR both equal u. Each class has probability one half, so when the whole budget is spent on one class, its cases can move by $2\varepsilon = s^2/2$ on average (Section 2.2). With a uniform density, moving an additional mass x across the threshold costs $x^2/2$, provided that this mass is available on the relevant side. Thus, if $s \le u \le 1-s$, the worst-case FPR is $u+s$ and the worst-case TPR is $u-s$. The same values follow from (2.10) and (2.11), with the mass closest to the threshold moved first. For the open event $\{m(X) > c\}$, the supremum is approached by moving mass to points just beyond the boundary, although it need not be attained. When $u \ge 1-s$, the worst-case FPR is one, and when $u \le s$, the worst-case TPR is zero. For thresholds with $\log\{c/(1-c)\}$ outside $[0,1]$, likewise, either the worst-case FPR is one or the worst-case TPR is zero.

For $\alpha \in [0.7, 0.9]$ and $s \in (0, 0.1)$, the smallest threshold that is feasible at level α has nominal rate $u = \alpha - s$, so

$$g_{P_0}(\alpha) = \alpha, \qquad g_\varepsilon(\alpha; P_0) = \alpha - 2s,$$

and $\Delta_\varepsilon(P_0) = 2s$. In this range, both worst-case rates have derivative $-1/\{c(1-c)\}$ with respect to c. Hence Assumption 1(ii) holds on any compact interval of thresholds whose worst-case FPRs range over a closed interval inside $(0.6, 1)$ that contains $[0.7, 0.9]$ in its interior. The score is logistic with a nonzero slope, and the relevant thresholds lie in a compact subinterval of $(0,1)$, so Lemma 3 gives Assumption 1(i). The doctor is robustly dominated, since $g_\varepsilon(0.8; P_0) - 0.2 = 0.6 - 2s > 0$. For every $Q \in \mathcal{B}_\varepsilon(P_0)$, $g_Q(0.8) \ge 0.8 - 2s$, so the indicator in (3.16) is zero for $t < 0.6 - 2s$ and is at most $t/(0.6-2s)$ for larger t. Assumption 4 therefore holds at every $Q \in \mathcal{B}_\varepsilon(P_0)$, with $\gamma_M(Q) = 1$ and $C_M(Q) = (0.6 - 2s)^{-1}$.

At $Q = P_0$, the doctor's welfare is $2(0.2) - 0.8 = -0.4$, so replacement improves on retention at both thresholds in Example 2.

A tail limit for Beta distributions. Suppose that $Z_N \sim \mathrm{Beta}(a_N, b_N)$ with $a_N/N \to a > 0$ and $b_N/N \to b > 0$, and let $\theta = a/(a+b)$. For fixed $x \in (\theta, 1)$,

$$-\frac{1}{N} \log \mathbb{P}(Z_N > x) \longrightarrow (a+b)\,\mathrm{kl}(\theta, x), \tag{A.22}$$

where kl is the function defined in the proof of Lemma 6. To see this, write the probability as the ratio of the integrals of the Beta density kernel over $[x,1]$ and $[0,1]$. The logarithm of the integrand, divided by N, is

$$f_N(z) = \frac{a_N - 1}{N} \log z + \frac{b_N - 1}{N} \log(1-z).$$

For large N, this function is strictly concave, its maximizer converges to θ, and it is decreasing on $[x,1]$. The numerator is therefore at most $\exp\{N f_N(x)\}$ and at least $h \exp\{N f_N(x+h)\}$ for $0 < h < 1 - x$; letting $N \to \infty$ and then $h \downarrow 0$ shows that its logarithm divided by N tends to $f(x) = a \log x + b \log(1-x)$. Bounding the denominator above by the value of the integrand at its mode and below by its integral over an interval around θ shows in the same way that the corresponding limit is $f(\theta)$. Subtracting gives (A.22).

The selected threshold. Let the N cases in the doctor's history be independent and identically distributed, with prevalence $1/2$, and let the doctor's true FPR and TPR be 0.8 and 0.2, as in Example 2. By the strong law of large numbers,

$$\frac{\mathrm{TN}_N}{N}, \frac{\mathrm{TP}_N}{N} \longrightarrow 0.1, \qquad \frac{\mathrm{FP}_N}{N}, \frac{\mathrm{FN}_N}{N} \longrightarrow 0.4 \quad \text{almost surely.}$$

Under the posterior, $Z_{0,N} := 1 - \alpha_H$ and $Z_{1,N} := \beta_H$ are independent Beta variables with shape parameters

$$(a_0 + \mathrm{TN}_N, a_0 + \mathrm{FP}_N), \qquad (a_0 + \mathrm{TP}_N, a_0 + \mathrm{FN}_N).$$

Let $F_{0,N}$ and $F_{1,N}$ be their CDFs. For $s < u < 1 - s$, the exact posterior probability (2.12) is

$$q_N(u) = F_{0,N}(1 - u - s)F_{1,N}(u - s).$$

This function is continuous and vanishes at the two endpoints, and the posterior probability is zero outside the interval, so a maximizer exists. Write $\widehat{u}_N$ for the nominal rate of the selected threshold. Put $x_0 = 0.5 - s > 0.2$. At $u = 0.5$,

$$1 - q_N(0.5) \le \mathbb{P}(Z_{0,N} > x_0) + \mathbb{P}(Z_{1,N} > x_0).$$

For any fixed $d \in (0, x_0 - 0.2)$ and all $u \ge 0.5 + d$,

$$1 - q_N(u) \ge \mathbb{P}(Z_{0,N} > x_0 - d),$$

and for $u \le 0.5 - d$, the same lower bound holds with $Z_{1,N}$ in place of $Z_{0,N}$. On the event of probability one on which the limits above hold, (A.22) applies to both variables with $(a, b) = (0.1, 0.4)$. Since $\mathrm{kl}(0.2, x)$ is strictly increasing in x for $x > 0.2$, the lower bounds for $|u - 0.5| \ge d$ are eventually strictly larger than the upper bound at $u = 0.5$, so every maximizer satisfies $|u - 0.5| < d$ for all sufficiently large N. Applying this to a sequence $d \downarrow 0$ proves that $\widehat{u}_N \to 0.5$ almost surely. Moreover, $q_N(0.5) \to 1$, so the doctor is eventually replaced for every fixed $\tau < 1$.

Because the posterior concentrates, the posterior median $a_{N,1}$ of the FPR also converges to 0.8. For $a_{N,1}$ close to 0.8, $g_\varepsilon(a_{N,1}; P_0) = a_{N,1} - 2s$, while the worst-case FPR and TPR at the selected threshold are $\widehat{u}_N + s$ and $\widehat{u}_N - s$. Consequently, the threshold term (3.17), which we write $\widehat{K}_N$ because only the history grows here, satisfies

$$\widehat{K}_N = \left[2\{a_{N,1} - 2s - (\widehat{u}_N - s)\} + (\widehat{u}_N + s - a_{N,1})\right]_+ \longrightarrow 0.3 - s.$$

The welfare of the selected threshold tends to 0.5, and that of the benchmark is 0.8, so $\mathrm{Regret}(\varepsilon, P_0) \to 0.3$, as claimed. Because the worst-case rates under P_0 are known, no part of this limit comes from estimation: s of it is the population regret, and the remaining $0.3 - s$ is the limit of the threshold term. □

A.3 Proofs for Section 4

The worst-case ROC curve under a pointwise budget.

Lemma 7 (Measurable local extrema)**.** *The selections in Definition 6 exist in either of the following cases:*

(i) $\mathcal{X}$ *is closed in* $\mathbb{R}^d$ *and* $m : \mathcal{X} \to [0,1]$ *is continuous;*

(ii) $\mathcal{X} = \mathbb{R}^d$ *and* m *is constant on each cell of a finite axis-aligned rectangular partition, whose open and closed faces are specified.*

A fixed finite ensemble of axis-aligned trees is of the form in (ii), with the partition obtained by intersecting the leaf cells of its trees.

Proof of Lemma 7. In case (i), $\mathbb{B}_\rho(x)$ is nonempty and compact, so the minimum and the maximum of m over it exist. For any real a, the set $\{x : \underline{m}_\rho(x) \le a\}$ is closed: if $x_n \to x$ and $z_n \in \mathbb{B}_\rho(x_n)$ satisfy $m(z_n) \le a$, then the z_n are bounded, and a subsequence converges to some $z \in \mathbb{B}_\rho(x)$ with $m(z) \le a$. Hence $\underline{m}_\rho$ is Borel measurable, and the same argument applied to $-m$ shows that $\overline{m}_\rho$ is as well.

A measurable selection can be constructed as follows. For a closed bounded cube D, let $v_D(x)$ be the minimum of m over $\mathbb{B}_\rho(x) \cap D$, with $v_D(x) = +\infty$ if this set is empty. The subsequence argument above shows that v_D is Borel measurable. The set of x for which some minimizer of m over $\mathbb{B}_\rho(x)$ lies in D is $\{x : v_D(x) = \underline{m}_\rho(x)\}$, which is therefore Borel. Enumerate closed unit cubes that cover $\mathbb{R}^d$ and, for each x, take the first one that contains a minimizer. Divide this cube into its 2^d closed subcubes of half the side length, take the first of these that contains a minimizer, and continue in this way. Each of these choices is measurable in x. The selected cubes are nested, each of them meets the compact set of minimizers, and their diameters tend to zero, so their centers converge to a minimizer, and the limit is measurable in x. The same construction applied to $-m$ gives the maximizer.

In case (ii), fix a nonempty cell D and a point $z_D \in D$, and let $\pi_D(x)$ denote the projection of x onto the closure of D, which is a closed rectangle; the map π_D is continuous. The cell D contains a point of $\mathbb{B}_\rho(x)$ if and only if

$$\|x - \pi_D(x)\| < \rho, \quad \text{or} \quad \|x - \pi_D(x)\| \le \rho \text{ and } \pi_D(x) \in D.$$

Indeed, the distance from x to a closed convex set is attained only at the projection of x. Under the strict inequality, the points $\pi_D(x) + 2^{-k}(z_D - \pi_D(x))$ lie in D and, for large k, in $\mathbb{B}_\rho(x)$. The set of x from which D can be reached is Borel. For such x, take $\pi_D(x)$ if it belongs to $D \cap \mathbb{B}_\rho(x)$, and otherwise the point above with the smallest k for which it belongs to $\mathbb{B}_\rho(x)$; this defines a measurable map from x to a point of $D \cap \mathbb{B}_\rho(x)$. At least one cell can be reached, because $x \in \mathbb{B}_\rho(x)$. Among the finitely many cells that can be reached, choose one with the smallest score, breaking ties by a fixed ordering of the cells, and likewise one with the largest score. The corresponding points are the required measurable selections. □

Proof of Proposition 3. For every map T_X as in Definition 5, we have, P-almost surely,

$$\underline{m}_\rho(X) \le m(T_X(X,Y)) \le \overline{m}_\rho(X).$$

Since the labels, and hence the prevalence, are unchanged, $\mathrm{TPR}_Q(c) \ge \tilde{\beta}_\rho(c; P)$ and $\mathrm{FPR}_Q(c) \le \tilde{\alpha}_\rho(c; P)$ for every $Q \in \tilde{\mathcal{B}}_\rho(P)$. The map that sends each sick case to $z_\rho^-(X)$ and each healthy case to $z_\rho^+(X)$ is measurable and defines a distribution in $\tilde{\mathcal{B}}_\rho(P)$. Because the selections attain the extrema, this distribution attains both bounds at every threshold, also for cases with $\underline{m}_\rho(X) = c$ or $\overline{m}_\rho(X) = c$. This proves (4.5) and (4.6).

At $\rho = 0$, we have $\mathbb{B}_0(x) = \{x\}$, so $\underline{m}_0 = \overline{m}_0 = m$ and $\tilde{g}_0(\cdot; P) = g(\cdot; P)$. For any $Q \in \tilde{\mathcal{B}}_\rho(P)$, the map T_X that defines Q gives a coupling of P and Q with cost at most ρ. Hence, if $\rho \le \varepsilon$,

$$\{P\} \subseteq \tilde{\mathcal{B}}_\rho(P) \subseteq \mathcal{B}_\varepsilon(P).$$

It follows that, at every threshold, the worst-case rates over $\tilde{\mathcal{B}}_\rho(P)$ lie between the nominal rates and the worst-case rates over $\mathcal{B}_\varepsilon(P)$, and taking, for each α, the supremum over the thresholds that are feasible at level α for each pair of rates gives (4.7). □

The requirement that the extrema be attained cannot be dropped for arbitrary Borel scores. On $\mathcal{X} = [0, 1]$, let $m(0) = 1$ and $m(z) = z$ for $z > 0$, and take $x = 0$, $\rho = 1$ and $c = 0$. Then $m(z) > 0$ for every $z \in \mathbb{B}_\rho(0)$, although $\underline{m}_\rho(0) = 0$. If all the sick cases lie at 0, then $\tilde{\beta}_\rho(0; P) = 0$, whereas $\mathrm{TPR}_Q(0) = 1$ for every $Q \in \tilde{\mathcal{B}}_\rho(P)$. Lemma 7 excludes such scores; in particular, the scores of finite tree ensembles attain their extrema although they are discontinuous.

Estimation, replacement and welfare. We use the notation of Section 4.2.

Proof of Theorem 5. We condition throughout on the independent training sample.

Part (i). Write $\pi_y = P_0(Y = y) > 0$ and $\widehat{\pi}_y = N_y/n$. The classes

$$\{\mathbb{I}(Y = y)\mathbb{I}(Z_y(X) \le c) : c \in [0, 1]\}, \qquad y = 0, 1,$$

consist of indicators of nested sets and have bounded envelopes and finite VC dimension. They are pointwise measurable, whether or not F_y has atoms, because the indicator at c is the pointwise limit of the indicators at rational thresholds decreasing to c. Their covering numbers are polynomial (van der Vaart and Wellner, 1996, Theorem 2.6.4), so the uniform entropy integral is finite, and the maximal inequality of van der Vaart and Wellner (2011, equation (1.1)) gives a uniform $O_p(n^{-1/2})$ bound for the centered empirical averages. Also, $\widehat{\pi}_y - \pi_y = O_p(n^{-1/2})$. On the event $\widehat{\pi}_y > 0$, the identity

$$\widehat{F}_y(c) - F_y(c) = \frac{(\widehat{\mathbb{P}}_n - P_0)[\mathbb{I}(Y = y)\mathbb{I}(Z_y(X) \le c)] - F_y(c)(\widehat{\pi}_y - \pi_y)}{\widehat{\pi}_y}$$

gives $d_{y,n} = O_p(n^{-1/2})$. Samples in which a class is absent have probability $p_0^n + (1 - p_0)^n$, and on them $\widehat{F}_y$ may be set equal to an arbitrary fixed CDF without affecting these rates.

The worst-case rates $1 - F_y$ under P_0 and their estimates $1 - \widehat{F}_y$ are right-continuous and nonincreasing in c. Since $c = 1$ is feasible at every level, the smallest threshold that is feasible at level α exists, and it has the highest worst-case TPR among the thresholds that are feasible at that level. A threshold that is feasible at level $\alpha - d_{0,n}$ under P_0 is feasible at level α under $\widehat{\mathbb{P}}_n$, and a threshold that is feasible at level α under $\widehat{\mathbb{P}}_n$ is feasible at level $\alpha + d_{0,n}$ under P_0; moreover, at every threshold, the worst-case TPR under P_0 and its estimate differ by at most $d_{1,n}$. These two inclusions and this bound prove (4.13).

Part (ii). By the endpoint inequalities in Assumption 5(ii), c_α lies in $\mathcal{C}$ for every α within b of I. For such α,

$$\tilde{g}_\rho(\alpha; P_0) = 1 - F_1(c_\alpha), \qquad \left|\frac{d}{d\alpha}\tilde{g}_\rho(\alpha; P_0)\right| = \frac{f_1(c_\alpha)}{f_0(c_\alpha)} \le L_1/\underline{f}.$$

Once $d_{0,n} < b$, (4.13) and this Lipschitz bound give an error of at most $(L_1/\underline{f})\, d_{0,n} + d_{1,n}$ for the estimated curve on I. By part (i), $d_{0,n} < b$ with probability tending to one, and this bound is $O_p(n^{-1/2})$, which proves the rate.

Part (iii). As noted before Theorem 5, every threshold gives, with probability one, the same indicators $\mathbb{I}\{Z_y(X) > c\}$ as one of $c_1, \ldots, c_M$, both under P_0 and for the validation cases. Let $d = \operatorname{dist}(I, \mathcal{L}) > 0$, and let $\widehat{A}_k$ and $\widehat{B}_k$ be the estimated worst-case FPR and TPR at c_k. Since $|\widehat{A}_k - A_k| \le d_{0,n}$ for every k, if $d_{0,n} < d/2$, then for every $\alpha \in I$ each c_k is feasible at level α under $\widehat{\mathbb{P}}_n$ exactly when it is feasible at level α under P_0. The index $k(\alpha)$ is then the same for the estimated curve and for $\tilde{g}_\rho(\cdot; P_0)$, and since the worst-case TPR is nonincreasing in the threshold,

$$\widehat{\tilde{g}}_\rho(\alpha) - \tilde{g}_\rho(\alpha; P_0) = \widehat{B}_{k(\alpha)} - B_{k(\alpha)}.$$

The supremum over $\alpha \in I$ of the absolute value of the right side is at most $d_{1,n}$, and, by part (i), $d_{0,n} < d/2$ with probability tending to one; this proves the rate. □

Differentiability under a pointwise budget. A direction for (F_0, F_1) is a pair (h_0, h_1) of functions that are continuous on $\mathcal{C}$. When F_0 and F_1 have finite supports, a direction for the probabilities that they put on their support points is a pair (v_0, v_1) of vectors $v_y = (v_{y,\ell})_\ell$, where $v_{y,\ell}$ is the change in the probability that F_y puts on s_ℓ.

Theorem 7 (Differentiability of the worst-case ROC curve under a pointwise budget)**.** *Suppose that Assumption 5(i) holds.*

(i) Under Assumption 5(ii), the map from (F_0, F_1) to $\tilde{g}_\rho(\cdot; P_0)$, with values in $\ell^\infty(I)$, is Hadamard differentiable in the uniform norm on $\mathcal{C}$, tangentially to the directions (h_0, h_1). Its derivative is

$$\dot{\tilde{g}}_{(h_0,h_1)}(\alpha) = -h_1(c_\alpha) + \frac{f_1(c_\alpha)}{f_0(c_\alpha)} h_0(c_\alpha). \tag{A.23}$$

(ii) Suppose instead that F_0 and F_1 have fixed finite supports, and let $I \subset (0,1)$ be compact with $\operatorname{dist}(I, \mathcal{L}) > 0$. As a function of the probabilities that F_0 and F_1 put on their support points, the map into $\ell^\infty(I)$ is Hadamard differentiable, with derivative

$$\dot{\tilde{g}}_{(v_0,v_1)}(\alpha) = \sum_{\ell : s_\ell > c_{k(\alpha)}} v_{1,\ell}. \tag{A.24}$$

Proof of Theorem 7. Part (i). Let $F_{y,t} = F_y + t h_{y,t}$ be CDFs with $\|h_{y,t} - h_y\|_{\infty,\mathcal{C}} \to 0$ as $t \downarrow 0$, where h_y is continuous on $\mathcal{C}$. For small t, the inverses at all $\alpha \in I$ lie in $\mathcal{C}$, and the feasibility inclusions used for (4.13), together with $f_0 \ge \underline{f}$, give

$$\sup_{\alpha \in I} |c_{\alpha,t} - c_\alpha| = O(t), \qquad c_{\alpha,t} = \inf\{c : F_{0,t}(c) \ge 1 - \alpha\}.$$

The perturbed CDF $F_{0,t}$ may have jumps, but $F_0 + t h_0$ is continuous on $\mathcal{C}$ and differs from $F_{0,t}$ by $o(t)$ uniformly on $\mathcal{C}$, so the jumps of $F_{0,t}$ at the points $c_{\alpha,t}$, $\alpha \in I$, are $o(t)$ uniformly in α. Since

$$F_{0,t}(c_{\alpha,t}-) \le 1 - \alpha \le F_{0,t}(c_{\alpha,t}),$$

it follows that $F_{0,t}(c_{\alpha,t}) - (1-\alpha) = o(t)$ uniformly in α. A Taylor expansion of F_0, uniform on $\mathcal{C}$, and the uniform continuity of h_0 then give

$$0 = f_0(c_\alpha)(c_{\alpha,t} - c_\alpha) + t h_0(c_\alpha) + o(t).$$

Thus $(c_{\alpha,t} - c_\alpha)/t \to -h_0(c_\alpha)/f_0(c_\alpha)$ uniformly in α, and expanding $1 - F_{1,t}(c_{\alpha,t})$ in the same way proves (A.23). The derivative is linear and continuous in (h_0, h_1), with norm at most $1 + L_1/\underline{f}$. Hence the map from (F_0, F_1) to $\tilde{g}_\rho(\cdot; P_0)$ is Hadamard differentiable with respect to the uniform norm on $\mathcal{C}$, tangentially to the directions (h_0, h_1). Differentiability as a function of P follows by composition with the map from P to (F_0, F_1), in which each F_y is normalized by π_y, as in the identity displayed in the proof of Theorem 5(i).

Part (ii). The argument in the proof of Theorem 5(iii) applies to any sufficiently small perturbation of the probabilities on the fixed supports. Each B_k is the sum of the probabilities that F_1 puts on the support points above c_k, so it is linear in them, and the last identity in that proof gives (A.24) uniformly on I. □

The condition $\operatorname{dist}(I, \mathcal{L}) > 0$ in Theorems 5(iii) and 7(ii) cannot be dropped, even with a positive budget, a single tree and exact local extrema. Consider the one-dimensional tree score $m(x) = 1/4$ for $x \le 0$ and $m(x) = 3/4$ for $x > 0$, with $\rho = 1/4$, and suppose that in each class X takes the values -1 and 1, with

$$P_0(X = 1 \mid Y = 0) = a, \qquad P_0(X = 1 \mid Y = 1) = b, \quad 0 < a, b < 1.$$

No ball $\mathbb{B}_\rho(x)$ centered at a support point crosses the split, so $\underline{m}_\rho = \overline{m}_\rho = m$ at both support points. The available pairs of rates are $(1, 1)$, (a, b) and $(0, 0)$, and

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

Let $\widehat{a}$ be the share of the healthy validation cases at 1. If $a = 1/2$, then under Assumption 5(i) the event $\{\widehat{a} > a\}$ has probability tending to $1/2$, and on this event the estimated curve is zero at a, whereas $\tilde{g}_\rho(a; P_0) = b$. Thus the estimated worst-case ROC curve is not even pointwise consistent at $\alpha = a$. Nor is the map from the probabilities on the support points to $\tilde{g}_\rho(a; P_0)$ directionally differentiable: changing the probability that a healthy case lies at 1 from a to $a + t$, for any $t > 0$, changes $\tilde{g}_\rho(a; P_0)$ from b to zero, although the two conditional CDFs change by at most t.

In the continuous case, it is likewise not enough that F_0 be strictly increasing. Let the CDFs of the worst-case scores on $[0, 1]$ be $F_0(c) = 1/2 + 4(c - 1/2)^3$ and $F_1(c) = c$. For any fixed $\rho > 0$, they arise from the score $m(x) = \sigma(x)$, with σ as in (4.9): if V_y has CDF F_y, let $X \mid Y = 0$ have the distribution of $\operatorname{logit}(V_0) - \rho$ and $X \mid Y = 1$ that of $\operatorname{logit}(V_1) + \rho$. Both CDFs are continuous and strictly increasing, but $f_0(1/2) = 0$. The CDF $F_{0,t}(c) = (1 - t)F_0(c) + t\mathbb{I}\{c \ge 1/4\}$ differs from F_0 by at most t, yet at $\alpha = 1/2$ its inverse differs from that of F_0 by $-[t/\{8(1-t)\}]^{1/3}$ for sufficiently small t, and the curve changes by the same amount with the opposite sign. At such a point, a uniform error of order $n^{-1/2}$ in the estimated CDFs therefore does not imply an error of order $n^{-1/2}$ in the curve. The lower bound on f_0 in Assumption 5(ii) rules this out.

Replacement and welfare under a pointwise budget.

Proof of Corollary 2. By Proposition 3, $\tilde{\alpha}_\rho$ and $\tilde{\beta}_\rho$ are the worst-case FPR and TPR over $\tilde{\mathcal{B}}_\rho(P_0)$, so (3.10) holds with the substituted objects for every $Q \in \tilde{\mathcal{B}}_\rho(P_0)$. Because the worst-case FPR is right-continuous and both rates are nonincreasing in the threshold, the smallest threshold that is feasible at level α attains $\tilde{g}_\rho(\alpha)$; in the proofs of Section 3, Lemma 2 provides this property for g_ε. For $Q \in \tilde{\mathcal{B}}_\rho(P_0)$, every threshold whose worst-case FPR is at most α has FPR under Q at most α and TPR under Q at least its worst-case TPR, so $\tilde{g}_\rho \le g_Q$. The arguments in Appendix A.2 then apply with the substituted objects under the stated conditions: the rectangles of the candidate thresholds in the proofs of Theorems 2 and 3, and the event $\mathcal{E}(t)$ and the decomposition (3.15) in the proof of Theorem 4. For the welfare choice, the argument at the end of Section 3.3 applies. □

A.4 Replacement using confidence bounds

If simultaneous confidence bounds on the worst-case rates are available, they can be used in place of the estimated worst-case rates in the rule of Definition 3. For each candidate $c \in \mathcal{C}_n$, let $U_{\alpha,n}(c) \in [0,1]$ be an upper confidence bound for the worst-case FPR $\overline{\alpha}_\varepsilon(c; P_0)$ and $L_{\beta,n}(c) \in [0,1]$ a lower confidence bound for the worst-case TPR $\underline{\beta}_\varepsilon(c; P_0)$, and suppose that the bounds are at least as conservative as the estimated worst-case rates:

$$U_{\alpha,n}(c) \ge \widehat{\overline{\alpha}}_\varepsilon(c), \qquad L_{\beta,n}(c) \le \widehat{\underline{\beta}}_\varepsilon(c). \tag{A.25}$$

An upper bound that violates (A.25) can be replaced by its maximum with $\widehat{\overline{\alpha}}_\varepsilon(c)$, and a lower bound by its minimum with $\widehat{\underline{\beta}}_\varepsilon(c)$; this does not reduce their coverage. Replacing the worst-case rates in (2.12) by the bounds gives, for each candidate c, the posterior probability

$$\widehat{q}_j^{\,L}(c) = \Pi_j\{\alpha_H \ge U_{\alpha,n}(c),\ \beta_H \le L_{\beta,n}(c) \mid \mathcal{D}_j\}. \tag{A.26}$$

We call the event in (A.26) the rectangle of c under the bounds. With the estimated worst-case rates in place of the bounds, it is the rectangle of c defined in the proof of Theorem 2.

The rule with confidence bounds replaces doctor j if a maximizer $\widehat{c}_j^{\,L}$ of (A.26) over $\mathcal{C}_n$ satisfies $\widehat{q}_j^{\,L}(\widehat{c}_j^{\,L}) \ge \tau$, and then uses the threshold $\widehat{c}_j^{\,L}$; otherwise it retains the doctor. As for the implemented rule, on a finite grid the maximizer exists and a fixed tie-breaking rule makes it measurable. Let

$$\mathcal{E}_n^L = \{\overline{\alpha}_\varepsilon(c; P_0) \le U_{\alpha,n}(c), \quad \underline{\beta}_\varepsilon(c; P_0) \ge L_{\beta,n}(c) \text{ for every } c \in \mathcal{C}_n\}$$

be the event that the bounds cover the worst-case rates at every candidate. For $\gamma \in (0,1)$, we say that the bounds have simultaneous coverage $1-\gamma$ if this event has probability at least $1-\gamma$.

Proposition 5 (Conditional guarantee for replacement using confidence bounds)**.** *Suppose that (2.4) and (A.25) hold. Then:*

(i) the set of doctors replaced by the rule with confidence bounds is contained in the set replaced by the implemented rule with the same candidates and the same cutoff τ;

(ii) on $\mathcal{E}_n^L$, for every doctor replaced by the rule with confidence bounds, the posterior probability that $\widehat{c}_j^{\,L}$ dominates the doctor under every $Q \in \mathcal{B}_\varepsilon(P_0)$ is at least τ;

(iii) if, in addition, Assumptions 2 and 3 hold,

$$\sup_{c\in\mathcal{C}_n}\big\{|U_{\alpha,n}(c)-\overline{\alpha}_\varepsilon(c;P_0)|+|L_{\beta,n}(c)-\underline{\beta}_\varepsilon(c;P_0)|\big\}=o_p(1),$$

and, for each dominated doctor and some t_0 as in Theorem 2, $\mathcal{C}_n$ contains candidates whose worst-case rates under P_0 converge to those of a threshold that is feasible at level $\alpha_{H,j}-t_0$ under P_0 and attains $g_\varepsilon(\alpha_{H,j}-t_0;P_0)$, then the probability that the rule with confidence bounds assigns at least one doctor the wrong replacement decision tends to zero as $n\to\infty$ and $n_{\min}\to\infty$.

For the implemented rule, Theorem 3 gives the counterpart of part (ii) only up to the estimation errors $e_{\alpha,n}$ and $e_{\beta,n}$, which are unobserved. The rule with confidence bounds removes these errors on $\mathcal{E}_n^L$, and by part (i), the cost is that it may retain doctors whom the implemented rule replaces. If the bounds have simultaneous coverage $1-\gamma$, part (ii) holds with probability at least $1-\gamma$ over the validation sample, whereas τ bounds a posterior probability given the doctor's history.

Proof of Proposition 5. By (A.25), for every candidate c and in every sample, the rectangle of c under the bounds is contained in the rectangle of c. Hence $\widehat{q}_j^L(c)\le\widehat{q}_j^\varepsilon(c)$ for every $c\in\mathcal{C}_n$, so that $\widehat{q}_j^L(\widehat{c}_j^L)\le\widehat{q}_{\max,j}^\varepsilon$, and every doctor replaced by the rule with confidence bounds is also replaced by the implemented rule. This proves (i).

On $\mathcal{E}_n^L$, every $\theta_H=(\alpha_H,\beta_H)$ in the rectangle of $\widehat{c}_j^L$ under the bounds satisfies

$$\mathrm{FPR}_Q(\widehat{c}_j^L)\le\overline{\alpha}_\varepsilon(\widehat{c}_j^L;P_0)\le U_{\alpha,n}(\widehat{c}_j^L)\le\alpha_H,$$

$$\mathrm{TPR}_Q(\widehat{c}_j^L)\ge\underline{\beta}_\varepsilon(\widehat{c}_j^L;P_0)\ge L_{\beta,n}(\widehat{c}_j^L)\ge\beta_H$$

for every $Q\in\mathcal{B}_\varepsilon(P_0)$, so that $\widehat{c}_j^L$ dominates a doctor with these error rates under every such Q. For a doctor replaced by the rule with confidence bounds, the posterior probability of this rectangle is $\widehat{q}_j^L(\widehat{c}_j^L)\ge\tau$, which proves (ii).

For (iii), let δ_n^L be the supremum in part (iii), so that $\delta_n^L=o_p(1)$. Consider first a dominated doctor, and let c_j° be the threshold in part (iii), which exists by Lemma 2. Its worst-case FPR is at most $\alpha_{H,j}-t_0$. Because $t_0\le d_H/2$, the level $\alpha_{H,j}-t_0$ lies in $[\alpha_l,\alpha_u]$, and Assumption 2 gives a worst-case TPR of at least $\beta_{H,j}+\delta_0-L_gt_0>\beta_{H,j}+t_0$. Let $c_n\in\mathcal{C}_n$ be the candidates in part (iii), and let a_n be the sum of the absolute differences between their worst-case rates and those of c_j°, which tends to zero. On the event $\{\delta_n^L+a_n\le t_0/2\}$, whose probability tends to one, $U_{\alpha,n}(c_n)\le\alpha_{H,j}-t_0/2$ and $L_{\beta,n}(c_n)>\beta_{H,j}+t_0/2$. The rectangle of c_n under the bounds then contains the neighborhood $\{\theta_H:\|\theta_H-\theta_{H,j}\|_\infty\le t_0/2\}$, so that $\widehat{q}_j^L(\widehat{c}_j^L)\ge\widehat{q}_j^L(c_n)\ge1-T_j(t_0/2)$. By Assumption 3 and Markov's inequality, $T_j(t_0/2)\to_p0$, so the doctor is replaced with probability tending to one.

Consider next a doctor who is not dominated, and fix $0<t<\min\{d_H/2,\delta_0/[4(1+L_g)]\}$. Since $\delta_n^L=o_p(1)$, the event that $\alpha_{H,j}+t+\delta_n^L\in[\alpha_l,\alpha_u]$ and $(1+L_g)(t+\delta_n^L)<\delta_0$ has probability tending to one. On this event, suppose that the rectangle of some candidate c under the bounds intersects the neighborhood $\{\theta_H:\|\theta_H-\theta_{H,j}\|_\infty\le t\}$. Then

$$\overline{\alpha}_\varepsilon(c;P_0)\le\alpha_{H,j}+t+\delta_n^L,\qquad\underline{\beta}_\varepsilon(c;P_0)\ge\beta_{H,j}-t-\delta_n^L.$$

Thus c is feasible at level $\alpha_{H,j}+t+\delta_n^L$ under P_0, and $\beta_{H,j}\le g_\varepsilon(\alpha_{H,j}+t+\delta_n^L;P_0)+t+\delta_n^L$. By the Lipschitz condition in Assumption 2(ii), the right side is less than $g_\varepsilon(\alpha_{H,j};P_0)+\delta_0$,

which contradicts Assumption 2(i). Hence, on this event, the rectangle of every candidate under the bounds lies in the complement of the neighborhood, whose posterior probability is $T_j(t)$. Since $T_j(t) \to_p 0$ for the same reason, $\max_{c \in \mathcal{C}_n} \widehat{q}_j^{L}(c) \to_p 0$, and the doctor is retained with probability tending to one.

By Assumption 2(i), every doctor falls under one of the two cases. Since J is fixed, the probability that at least one doctor is assigned the wrong replacement decision tends to zero. □

B Details of the Monte Carlo simulations

Under (5.1) and (5.2), the error rates of the doctors and of the algorithm reduce to one-dimensional integrals. The variable R in (5.1) is normal with mean ξ_0 and variance $v = \|\xi\|_2^2 + \sigma_u^2$. With $\sigma(r) = (1+e^{-r})^{-1}$ the logistic function of Section 4.1, $p(Z, U) = \sigma(R)$, and the prevalence is $p_0 = \mathbb{E}[\sigma(R)]$. Because Y depends on (Z, U) only through R, and ν_j is independent of (Z, U, Y), the label is independent of (Z, U, ν_j) given R. With Φ the standard normal CDF, the probability that $\widehat{Y}_j = 1$ given $R = r$ is

$$H_j(r) = \Phi\left(\frac{\sigma(r) - c_{H,j}}{\sigma_\nu}\right).$$

Doctor j's error rates, which do not depend on δ, are therefore

$$\alpha_{H,j} = \frac{\mathbb{E}[(1-\sigma(R))H_j(R)]}{1-p_0}, \qquad \beta_{H,j} = \frac{\mathbb{E}[\sigma(R)H_j(R)]}{p_0}. \tag{B.1}$$

Let $m(x) = \{1 + \exp[-(w_0 + x^\top w)]\}^{-1}$ be the fitted score in a replication, with $w \neq 0$. We compute its error rates with (w_0, w) held fixed. Let $M = w_0 + Z^\top w$, the logit of $m(Z)$, and let w_1 be the first coordinate of w. Since (M, R) is jointly normal,

$$M \mid R = r \sim N(\mu(r), s_M^2), \qquad \mu(r) = w_0 + \frac{w^\top \xi}{v}(r - \xi_0), \qquad s_M^2 = \|w\|_2^2 - \frac{(w^\top \xi)^2}{v} > 0.$$

The inequality $s_M^2 > 0$ follows from the Cauchy–Schwarz inequality, because $w \neq 0$ and $\sigma_u > 0$. For a threshold $c \in (0, 1)$, let $t = \log\{c/(1-c)\}$, so that $m(x) > c$ if and only if $w_0 + x^\top w > t$. Under Q_δ, the algorithm sees X_δ in (5.2), and $w_0 + X_\delta^\top w = M + (1-2Y)\delta w_1$. Since M is a function of Z, it is independent of Y given R, so

$$\begin{aligned} \mathrm{FPR}_{Q_\delta}(c) &= \frac{\mathbb{E}[(1-\sigma(R))\Phi\{(\mu(R) + \delta w_1 - t)/s_M\}]}{1-p_0}, \\ \mathrm{TPR}_{Q_\delta}(c) &= \frac{\mathbb{E}[\sigma(R)\Phi\{(\mu(R) - \delta w_1 - t)/s_M\}]}{p_0}. \end{aligned} \tag{B.2}$$

At $c = 0$, both rates are one, and at $c = 1$, both are zero.

Let $t_{0,j}$ solve $\mathbb{P}(M > t_{0,j} \mid Y = 0) = \alpha_{H,j}$. Because $s_M > 0$ and $\sigma(R) < 1$, the distribution of M given $Y = 0$ has a positive density on the real line, and $0 < \alpha_{H,j} < 1$ by (B.1), so this solution exists and is unique. Under Q_δ, the FPR at a threshold c is $\mathbb{P}(M > t - \delta w_1 \mid Y = 0)$, so the threshold at which the algorithm has doctor j's FPR is the c with $\log\{c/(1-c)\} = t_{0,j} + \delta w_1$. Its TPR is $\mathbb{P}(M > t_{0,j} + 2\delta w_1 \mid Y = 1)$, and since both rates decrease in the threshold, this TPR equals $g_{Q_\delta}(\alpha_{H,j})$. The matched-FPR

benchmark is therefore defined at every Q_δ: it replaces doctor j if $\beta_{H,j}$ is below this TPR, and then uses this threshold. The rules, by contrast, use the same threshold for a replaced doctor at every δ: the selected threshold $\widehat{c}^q$ or, in the comparison of Section 5.5, the welfare choice $\widehat{c}^W$. The error rates of this threshold under Q_δ follow from (B.2).

We compute the integrals in (B.1) and (B.2), which are expectations over the normal distribution of R, by Gauss–Hermite quadrature with 1,024 nodes. With 512 nodes instead, the doctors' error rates and the algorithm's error rates at the thresholds used by the rules and at the matched-FPR thresholds change by at most 5×10^{-7}. If a rule replaces nobody in a replication, the share of harmful replacements among all doctors is zero in that replication, but the share among the replaced doctors is not defined, and its mean is taken over the replications with replacements.

The data supplement reports the results with the selected threshold at every combination of budget, deployment distance and welfare weights, including the 147 combinations for the finer grid of average budgets in Appendix B.2. It also reports the paired differences from the nominal rule and between the two threshold choices, the share of harmful replacements among all doctors and the sizes of the losses, including the losses averaged over all doctors. For each mean taken over the replications with replacements, it gives the number of these replications. It also contains figures of the evaluations of Section 5.3 at $\delta = 0$, $\varepsilon/2$ and ε under all three pairs of welfare weights, and of the evaluations of Section 5.4 under the maps T_a with $a = 0$, $\rho/2$ and ρ.

B.1 Alternative welfare weights

The replacement decisions and the selected thresholds are the same under all three pairs of welfare weights (Section 5.1). Table B.1 reports the paired difference between the welfare of the robust rule and that of the nominal rule at each deployment distance. With every positive average budget in $\{0.05, 0.1, 0.2, 0.3\}$, the robust rule retains all doctors (Section 5.2), so each entry applies to all four budgets. Under each pair of weights, the difference is negative at every deployment distance up to $\delta = 0.1$ and positive at $\delta = 0.3$, 0.5 and 1.

Table B.1: Robust-minus-nominal welfare under the prespecified weights.

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)

Notes: Means over the 100 replications, with Monte Carlo standard errors in parentheses; the differences are paired within replications. The entries are the same for every positive average budget in $\{0.05, 0.1, 0.2, 0.3\}$. Positive entries favor the robust rule, which retains all doctors. The column for $(2, 1)$ repeats the robust-minus-nominal column of Table 1.

B.2 Smaller average budgets

Figure B.1 shows the estimated worst-case ROC curves of the first replication for the seven average budgets of Section 5.3. Table B.2 reports the values plotted in Figure 4, together with the share of doctors replaced and Monte Carlo standard errors. Table B.3 reports the evaluations at $\delta = \varepsilon$ of Table 2 under the other two pairs of welfare weights.

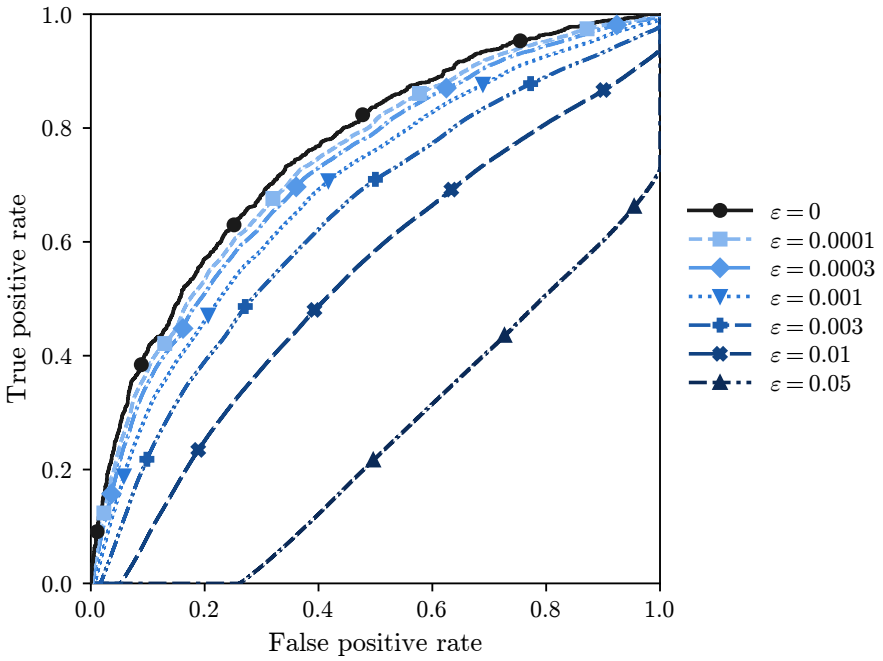


Figure B.1: Worst-case ROC curves for the seven average budgets.

Notes: Estimated worst-case ROC curves for the first replication in Section 5, using its fitted score, validation cases and 1,001 candidate thresholds; with $\varepsilon = 0$, the curve is the estimated nominal ROC curve. Each curve gives the highest worst-case TPR subject to the limit on the worst-case FPR on the horizontal axis. Steps correspond to candidate thresholds.

Table B.2: Share of doctors replaced and gain over retention, by average budget.

		Gain over retention, $(\phi, \eta) = (2, 1)$						
ε	Replaced (%)	$\delta = 0$	0.025	0.05	0.1	0.3	0.5	1
0	41.51 (0.88)	0.08257 (0.00173)	0.07274 (0.00153)	0.06284 (0.00133)	0.04283 (0.00093)	-0.03923 (0.00098)	-0.12279 (0.00266)	-0.32258 (0.00676)
0.0001	22.79 (0.70)	0.04628 (0.00138)	0.04083 (0.00122)	0.03532 (0.00105)	0.02419 (0.00073)	-0.02166 (0.00074)	-0.06860 (0.00214)	-0.18150 (0.00550)
0.0003	12.98 (0.49)	0.02666 (0.00099)	0.02354 (0.00087)	0.02039 (0.00076)	0.01401 (0.00052)	-0.01233 (0.00051)	-0.03940 (0.00151)	-0.10477 (0.00394)
0.001	3.18 (0.17)	0.00659 (0.00035)	0.00582 (0.00031)	0.00505 (0.00027)	0.00347 (0.00018)	-0.00307 (0.00018)	-0.00983 (0.00054)	-0.02627 (0.00143)
0.003	0.14 (0.02)	0.00028 (0.00004)	0.00025 (0.00003)	0.00022 (0.00003)	0.00015 (0.00002)	-0.00013 (0.00002)	-0.00043 (0.00006)	-0.00115 (0.00016)
0.01	0.00 (0.00)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)
0.05	0.00 (0.00)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)

Notes: Means over the 100 replications of Table 1, with Monte Carlo standard errors in parentheses; the gains over retention are paired within replications. At $\varepsilon = 0.003$, at least one doctor is replaced in 44 replications, but the means are taken over all 100. For the average budgets from 0.0001 to 0.01, $\delta = 0$ is the only deployment distance in the table at which Q_δ lies in $\mathcal{B}_\varepsilon(P_0)$; Tables 2 and B.3 report evaluations at $\delta = \varepsilon$.

Table B.3: Small average budgets at the boundary of their sets, other welfare weights.

	Robust rule		Nominal rule	
ε	Gain over retention	Harmful (%)	Gain over retention	Robust minus nominal
A. $(\phi, \eta) = (1, 1)$				
0.0001	0.02882	0.03	0.05071	−0.02189
	(0.00085)	(0.02)	(0.00104)	(0.00029)
0.0003	0.01675	0.02	0.05066	−0.03390
	(0.00062)	(0.02)	(0.00104)	(0.00051)
0.001	0.00424	0.00	0.05048	−0.04624
	(0.00023)	(0.00)	(0.00103)	(0.00085)
0.003	0.00019	0.00†	0.04998	−0.04980
	(0.00003)	(0.00)	(0.00102)	(0.00101)
0.01	0.00000	–	0.04824	−0.04824
	(0.00000)		(0.00099)	(0.00099)
0.05	0.00000	–	0.03818	−0.03818
	(0.00000)		(0.00078)	(0.00078)
B. $(\phi, \eta) = (5, 1)$				
0.0001	0.09859	0.34	0.17801	−0.07942
	(0.00300)	(0.07)	(0.00386)	(0.00126)
0.0003	0.05624	0.21	0.17784	−0.12160
	(0.00211)	(0.07)	(0.00386)	(0.00209)
0.001	0.01352	0.08	0.17727	−0.16375
	(0.00073)	(0.08)	(0.00384)	(0.00330)
0.003	0.00055	0.00†	0.17563	−0.17508
	(0.00007)	(0.00)	(0.00381)	(0.00379)
0.01	0.00000	–	0.16987	−0.16987
	(0.00000)		(0.00370)	(0.00370)
0.05	0.00000	–	0.13680	−0.13680
	(0.00000)		(0.00304)	(0.00304)

Notes: As in Table 2, for the welfare weights $(1, 1)$ and $(5, 1)$. The replacement decisions, and hence the shares of doctors replaced, are those of Table 2. A dagger marks a mean over the 44 replications with replacements.

B.3 Pointwise budgets and threshold choices

Tables B.4 and B.5 report the welfare and the share of harmful replacements for the two threshold choices, at each of the five pointwise budgets and seven deployment distances. With the prespecified positive average budgets, nobody is replaced under either threshold choice, and with a budget of zero, both kinds of budget give the nominal rule. Table B.6 reports the gains under the maps T_a of Section 5.4.

Table B.4: Welfare under a pointwise budget for the two threshold choices.

$\rho\backslash\delta$	0	0.025	0.05	0.1	0.3	0.5	1
A. Selected threshold $\widehat{c}^q$: gain over retention							
0	0.08257†	0.07274	0.06284	0.04283	-0.03923	-0.12279	-0.32258
	(0.00173)	(0.00153)	(0.00133)	(0.00093)	(0.00098)	(0.00266)	(0.00676)
0.05	0.03642†	0.03196†	0.02748†	0.01841	-0.01890	-0.05701	-0.14851
	(0.00123)	(0.00108)	(0.00093)	(0.00063)	(0.00071)	(0.00200)	(0.00509)
0.1	0.01063†	0.00929†	0.00794†	0.00520†	-0.00606	-0.01760	-0.04543
	(0.00053)	(0.00047)	(0.00040)	(0.00026)	(0.00032)	(0.00090)	(0.00230)
0.2	0.00025†	0.00021†	0.00018†	0.00010†	-0.00019	-0.00049	-0.00122
	(0.00003)	(0.00003)	(0.00002)	(0.00002)	(0.00003)	(0.00007)	(0.00017)
0.3	0.00000†	0.00000†	0.00000†	0.00000†	0.00000†	0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
B. Welfare choice $\widehat{c}^W$: difference from $\widehat{c}^q$							
0	0.01196†	0.01193	0.01191	0.01183	0.01123	0.01023	0.00646
	(0.00037)	(0.00037)	(0.00036)	(0.00036)	(0.00033)	(0.00029)	(0.00017)
0.05	0.00416†	0.00415†	0.00415†	0.00413	0.00397	0.00366	0.00243
	(0.00018)	(0.00018)	(0.00018)	(0.00018)	(0.00016)	(0.00015)	(0.00009)
0.1	0.00101†	0.00101†	0.00101†	0.00101†	0.00098	0.00092	0.00064
	(0.00006)	(0.00006)	(0.00006)	(0.00006)	(0.00006)	(0.00005)	(0.00003)
0.2	0.00002†	0.00002†	0.00002†	0.00002†	0.00002	0.00002	0.00001
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
0.3	0.00000†	0.00000†	0.00000†	0.00000†	0.00000†	0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)

Notes: Means over the 100 replications, including those without replacements, with Monte Carlo standard errors in parentheses; differences are paired within replications. The welfare weights are $(2, 1)$. A dagger marks the pairs (ρ, δ) at which Q_δ lies in $\tilde{\mathcal{B}}_\rho(P_0)$, that is, $\delta \leq \rho$. Standard errors shown as 0.00000 are below 0.000005. The row $\rho = 0$ is the nominal rule. With the prespecified positive average budgets, nobody is replaced, so the gain over retention and the difference between the two threshold choices are zero.

Table B.5: Share of harmful replacements under a pointwise budget for the two threshold choices.

$\rho\backslash\delta$	0	0.025	0.05	0.1	0.3	0.5	1
A. Selected threshold $\widehat{c}^q$							
0	0.17†	0.31	0.70	2.82	99.94	100.00	100.00
	(0.03)	(0.05)	(0.08)	(0.20)	(0.02)	(0.00)	(0.00)
0.05	0.24†	0.40†	0.74†	2.99	99.98	100.00	100.00
	(0.06)	(0.09)	(0.13)	(0.27)	(0.02)	(0.00)	(0.00)
0.1	0.13†	0.25†	0.65†	4.04†	100.00	100.00	100.00
	(0.06)	(0.08)	(0.14)	(0.59)	(0.00)	(0.00)	(0.00)
0.2	0.00†	0.00†	3.37†	11.52†	100.00	100.00	100.00
	(0.00)	(0.00)	(2.28)	(4.37)	(0.00)	(0.00)	(0.00)
0.3	–	–	–	–	–	–	–
	–	–	–	–	–	–	–
B. Welfare choice $\widehat{c}^W$							
0	0.03†	0.06	0.17	0.86	98.71	100.00	100.00
	(0.01)	(0.02)	(0.03)	(0.09)	(0.13)	(0.00)	(0.00)
0.05	0.10†	0.12†	0.31†	1.11	99.66	100.00	100.00
	(0.03)	(0.04)	(0.07)	(0.14)	(0.06)	(0.00)	(0.00)
0.1	0.04†	0.07†	0.16†	1.60†	99.92	100.00	100.00
	(0.04)	(0.05)	(0.08)	(0.31)	(0.04)	(0.00)	(0.00)
0.2	0.00†	0.00†	2.13†	8.69†	100.00	100.00	100.00
	(0.00)	(0.00)	(2.13)	(3.77)	(0.00)	(0.00)	(0.00)
0.3	–	–	–	–	–	–	–
	–	–	–	–	–	–	–

Notes: Mean share of harmful replacements among the replaced doctors, in percent, with Monte Carlo standard errors in parentheses; welfare weights and daggers as in Table B.4. The means are over all 100 replications for $\rho = 0$, 0.05 and 0.1 and over the 47 replications with replacements for $\rho = 0.2$; with $\rho = 0.3$, nobody is replaced, and the share is not defined.

Table B.6: Gain over retention under the maps T_a.

	$a = \rho/2$			$a = \rho$		
ρ	$\widehat{c}^q$	$\widehat{c}^W$	Nominal	$\widehat{c}^q$	$\widehat{c}^W$	Nominal
0	0.08257	0.09453	0.08257	0.08257	0.09453	0.08257
	(0.00173)	(0.00207)	(0.00173)	(0.00173)	(0.00207)	(0.00173)
0.05	0.03116	0.03532	0.07097	0.02586	0.03001	0.05927
	(0.00105)	(0.00122)	(0.00149)	(0.00087)	(0.00104)	(0.00125)
0.1	0.00745	0.00846	0.05927	0.00421	0.00522	0.03559
	(0.00038)	(0.00043)	(0.00125)	(0.00022)	(0.00027)	(0.00079)
0.2	0.00008	0.00010	0.03559	−0.00009	−0.00008	−0.01264
	(0.00001)	(0.00002)	(0.00079)	(0.00002)	(0.00001)	(0.00048)
0.3	0.00000	0.00000	0.01160	0.00000	0.00000	−0.06167
	(0.00000)	(0.00000)	(0.00042)	(0.00000)	(0.00000)	(0.00138)

Notes: Mean gain over retention over the 100 replications, with Monte Carlo standard errors in parentheses, $(\phi, \eta) = (2, 1)$. The rules with the pointwise budget ρ and the nominal rule are evaluated under the distribution of $(T_a(X, Y), Y)$, with T_a as in Section 5.4. With $a = \rho/2$, this distribution lies inside $\tilde{\mathcal{B}}_\rho(P_0)$; with $a = \rho$, it is the map (4.12), which attains both worst-case rates. The row $\rho = 0$ is the nominal rule under P_0. The two threshold choices replace the same doctors, and the replacement decisions do not depend on a.

C Details of the income-data illustration

C.1 Sample construction and evaluation

We use the 2018 one-year ACS data and the ten variables of the ACSIncome task: age, class of worker, educational attainment, marital status, occupation, place of birth, relationship to the householder, usual weekly hours worked, sex and race. Except for age and hours worked, these variables are categorical and enter X as numeric codes. We apply the eligibility restrictions of the task, namely age above 16, income above \$100, positive hours worked and a person weight of at least one, and give all eligible cases equal weight. From each of the three training states, we draw a fixed sample of 20,000 eligible cases. Of these 60,000 cases, 48,000 are used to fit the two scores, 6,000 form the validation sample, and the remaining 6,000 are the histories of the decision makers. None of the ten variables is missing for any eligible case in the eight states we use. Each variable is standardized by its mean and standard deviation in the 48,000 training cases.

The same 200 synthetic decision makers are used throughout, with thresholds $c_{H,j}$ drawn from $0.15 + 0.7\,\mathrm{Beta}(2,3)$, as in Section 5. In the histories and in each evaluation state, the cases are assigned at random to the decision makers, evenly within each class, so that every decision maker has cases of both classes. To generate decision maker j's decision on a case, we take the logit of the probability of high income given by the LR, add $U \sim N(0, 0.15^2)$ and apply the logistic function, which gives a probability $\widetilde{p}$. The decision is $\mathbb{I}\{\widetilde{p} + \nu_j > c_{H,j}\}$, with $\nu_j \sim N(0, 0.225^2)$, and U and ν_j are drawn for each decision, independently of the labels. The decisions on the history cases give each decision maker's confusion matrix and hence the Beta posteriors (2.17).

For the longer histories of Table 5, we draw further eligible cases without replacement from the cases of the three training states that are not used elsewhere. Of each decision maker's 150 cases, 59 or 60 have high income, and of the 300 cases, 118 or 119, about the same share as in the original histories. Within each class, the added cases come from the three states in the same proportions as in the original histories and are assigned to the decision makers at random. Each decision maker's history of 300 cases contains that of 150 cases, which contains the original history. The decisions on the added cases are generated in the same way. The rules with the added budgets are computed as above: the LR with a pointwise budget uses the exact local extrema (4.9), and for the RF, the budgets 0.01, 0.02 and 0.05 use one new search with $K = 200$ proposals, rescaled to each budget, while the larger budgets reuse the search of Section 6.

For both scores and all budgets, the candidate thresholds are 0, 1 and 400 equally spaced values from 0.001 to 0.999. At $\rho = 0.5$, the RF searches with $K \in \{20, 50, 100, 200, 500, 1000\}$ use the first K proposals of one sequence of 1,000 proposals, so that each search contains the smaller ones; with $K = 200$, this is the search used in Section 6. Since the categorical variables enter as numeric codes, the sets also allow these codes to move, although such moves need not correspond to any change in personal characteristics that could actually occur. For these variables, the worst-case rates show how sensitive the scores are to the codes.

In each evaluation state, we use all eligible cases, up to 20,000. Welfare is (3.12) with $(\phi, \eta) = (2, 1)$, and its error rates are computed from these cases. For a retained decision maker, they are the rates of the decision maker's own decisions on the assigned cases; for a replaced decision maker, they are the rates of the algorithm's decisions at the selected threshold on the same cases. The LR and the RF are evaluated on the same cases and

the same decisions.

The AUCs in Table 4 and in Panel C of Table C.1 are the conventional areas under the ROC curve of each score, computed from the cases on which the rules are evaluated. Unlike (2.2), the conventional ROC curve fills the gaps between the points reached by single thresholds by choosing at random between two thresholds (Section 2.1). The areas in Panels A and B of Table C.1 are areas under the estimated worst-case ROC curves over the candidate thresholds.

The age shift uses 5,000 further California cases that are not among the 60,000 cases above. These cases are assigned to the decision makers as in an evaluation state. We add δ to the standardized age seen by the algorithm, so δ is measured in standard deviations of age in the training cases. The labels stay fixed, and so do the decisions, which are generated from the unshifted cases. The nominal rules and the rules with budget 0.5 are those computed from the training states. By the argument that gives (5.3), the distance $\mathcal{W}_\kappa$ between the distributions of the shifted and the unshifted cases is δ.

C.2 Results by budget, number of proposals and age shift

Table C.1 reports the results by budget, number of proposals and age shift. The data supplement reports the results for every rule in each evaluation state and at each age shift, the confusion matrix of each decision maker, welfare computed from all cases of each state together rather than from each decision maker's cases, and the shares of harmful replacements among the replaced decision makers. These shares are computed from the cases on which the rules are evaluated and are not defined when nobody is replaced.

Table C.1: Results by budget, number of proposals and age shift in the income data.

A. By budget

	Area		Largest posterior probability	
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 with $\rho = 0.5$, by number of proposals

K	Area	Largest posterior probability	Number 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

C. Age shift in the further California cases: nominal rules

	Gain over retention		
δ	LR	RF	AUC of the RF
0	+0.00750	+0.01316	0.8814
0.5	+0.00845	+0.01227	0.8737
1	+0.01212	+0.01360	0.8687
1.5	+0.01100	+0.01424	0.8633
2	+0.00860	+0.01258	0.8609
2.5	+0.00883	+0.01055	0.8569
3	+0.01048	+0.01022	0.8508

Notes: In Panel A, the budget is an average budget for the LR, with the estimated worst-case rates computed exactly, and a pointwise budget for the RF, with a search over $K = 200$ proposals; budget 0 gives the nominal rule. The areas are those under the estimated worst-case ROC curves (Appendix C.1), and the largest posterior probability of dominance is taken over the 200 decision makers. Panel B keeps the fitted RF, the validation sample and the sequence of proposals fixed; its row $K = 200$ repeats the RF entries of Panel A at budget 0.5. Panel C uses the same 5,000 cases and decisions at every δ. The welfare of retaining all decision makers is 1.13206 with $(\phi, \eta) = (2, 1)$, and the AUC of the LR is 0.85690 at every δ, because the shift adds the same constant to the logit of the LR for every case. The rules with budget 0.5 retain all decision makers, so their gain over retention is zero at every δ. All results come from a single sample split, and no standard errors are reported.