Artificial-Intelligence Assisted Decision Making: A Statistical Framework

Han Hong* Xin Lin† Ke Tang‡ Jingyuan Wang§

December 22, 2019

Abstract. This paper proposes a statistical framework in which artificial intelligence can assist human decision making. Using observational data we benchmark the performance of each decision maker against the machine predictions, and replace decision makers whose information process quality is dominated by machine predictions based on the proposed criteria. The statistical frameworks that we proposed are applicable based on both Bayesian principles and frequentist principles of hypothesis testing and confidence set formation. Our theoretical discussion is illustrated by an example of birth defect detection, using a large data set of pregnancy outcomes and doctor diagnosis from the Pre-Pregnancy Checkups of reproductive age couples that are provided by the Chinese Ministry of Health. Based on doctor's diagnosis, we find doctors, especially those who are from rural areas, can be replaced by the machine learning prediction. Statistically, the overall quality of our algorithm on a testable data set outperforms the diagnoses made only by doctors, with higher true positive rate and lower false positive rate. Our example also informs that decision making with artificial intelligence is more beneficial to poor areas relative to developed places.

Keywords: artificial intelligence, machine learning, decision making, ROC curve

*Department of Economics, Stanford University. doubleh@stanford.edu
†Department of Computer Science, Beihang University. sweeneylin@buaa.edu.cn
‡Institute of Economics, School of Social Sciences, Tsinghua University.ketang@tsinghua.edu.cn
§Department of Computer Science, Beihang University. jywang@buaa.edu.cn

Electronic copy available at: https://ssrn.com/abstract=3508224

1 Introduction

In the era of machine learning, novel Artificial-Intelligence (AI) algorithms are capable of learning important features from both specialists and a large amount of data, and assisting in decision making in many scientific and engineering disciplines. For example, Gulshan et al. (2016) use a deep learning algorithm to detect diabetic retinopathy in retinal fundus photographs. Rudin et al. (2012) develop a machine learning model to assist power companies in predicting potential failures of components and systems. Babylon Health develops an AI based online diagnosis system that can provide triage advice to patients, and experiments show that their system has comparable accuracy with human doctors (Razzaki et al. (2018)). Rasch et al. (2003) use AI technology to build decision support system in military. Liu et al. (2017) use convolutional neutral network to help doctors detecting cancer metastases. Castro et al. (2017) use natural language process (NLP) techniques to identify patients with cerebral aneurysms from electronic medical record (EMR). Aletras et al. (2016) use NLP techniques to build models that can learn to predict the court's judgment based on textual content from corresponding cases. Currie and MacLeod (2017) analyzes the decision making of physicians, and informs the possibility of improvements that can both benefit patients and decrease medical expenses.

Much of the literature compares the performance of machines with the representative (average) person. For example, Rajpurkar et al. (2017) trained a 34-layer convolutional neural network to process ECG sequences and compared its performance to 6 cardiologists. Esteva et al. (2017) proposed a deep convolutional neural network structure for skin cancer classification and claimed that the model outperforms the average dermatologist. Kermany et al. (2018) proposed an image-based deep learning model to classify macular degeneration and concluded that it outperforms human. The conclusions of these papers are mostly based on observing an average pair of true positive and false positive rates that lie strictly below the Receiver Operating Characteristic (ROC) curve formed by the machine classification algorithm, implying that machines can achieve a higher True Positive Rate (TPR) for a given False Positive Rate (FPR), or a lower FPR for a given TPR.[1] However, an important message of our current paper is to *caution against* such interpretations without a deeper understanding of the human decision making process. Firstly, such findings can be rationalized not only by the superior information quality of machine learning algorithms, but also by the *incentive heterogeneity* of human decision makers who can be as intelligent as machine learning algorithms in processing statistical information from observational data.

[1] FPR is the number of negative events with wrong classification divided by the number of actual negative events, and TPR is the number of positive events with correct classification divided by the number of total positive events. A pair of FPR/TPR depicts the tradeoff between size and power for classical hypothesis tests.

Electronic copy available at: https://ssrn.com/abstract=3508224

Secondly, the FPR/TPR pairs of human decision makers are *not precisely measured*, especially when the number of cases that the decision makers see is not large. The example in our paper clearly shows strong heterogeneity in human decision makers; it also presents that the randomness in measuring decision makers play a key role in comparing the performance of human decision makers and machine algorithms.

To illustrate the first issue of concern, consider Figure 1, in which a collection of human decision makers, denoted $j = 1, \ldots, J$, all lie approximately on the machine-learned ROC. This is the case if they employed decision rules $\hat{Y}_i = \mathbb{1}\left(\hat{p}\left(X_i\right) > c_i\right)$ with the same $\hat{p}\left(\cdot\right)$ function but with different individual cutoff points c_i. Yet, after aggregating over all decision makers, the aggregate TPR/FPR pair lies visibly below the ROC.

[Figure 1 about here.]

This is an immediate artifact of Jensen's inequality due to the concavity of the observed ROC, and bears no implication on the comparison between the qualities of the machine learning algorithm and human decision makers. As shown in Feng et al. (2019), an optimal ROC is necessarily concave. This simple observation appears to have gone largely unnoticed by the literature.

More precisely, as long as the collection of humans' individual TPR/FPR points can be represented by a concave curve, the aggregated humans' TPR/FPR must fall below the curve of humans' individual TPR/FPR points. Let α_j denote FPR and β_j denote TPR, and suppose they are related by $\beta = f\left(\alpha\right)$, where $f\left(\cdot\right)$ is concave. Then by Jensen's inequality:

$$\bar{\beta} = \frac{1}{J}\sum_{j=1}^{J}\beta_j = \frac{1}{J}\sum_{j=1}^{n} f\left(\alpha_j\right) < f\left(\frac{1}{J}\sum_{j=1}^{J}\alpha_i\right) = f\left(\bar{\alpha}\right).$$

The foregoing discussion highlights the need to present a framework to statistically compare the performance of machine algorithms and human decision.

A machines binary classification rule is typically based on $D = \mathbb{1}(p(x) > c)$, where x is a set of features, the propensity score function $p\left(x\right)$ is learned from a large amount of data, and c is a cutoff threshold. The entire ROC curve will be generated across various cutoff threshold c. In the meanwhile, using historical data, we can compute a fixed FPR/TPR pair for each individual decision maker.

With assumptions 1 and 2 in Section 2, we can compare the machine ROC and the individual TPR/FPR pair when it is precisely known without sampling errors. For example, if human's TPR/FPR pair is precisely known below the machine ROC curve, we can find a collection of points on the machine's ROC curve which has a larger TPR and/or smaller FPR than the human decision maker. As shown in Figure 2, these correspond to the fraction on the ROC curve between A and B.

Electronic copy available at: https://ssrn.com/abstract=3508224

However, in reality only a *noisy* estimate of the TPR/FPR pair can be obtained from empirical data. This paper directly address the issue of accounting for sampling errors in making such a comparison. In particular, we address two important questions. The first is when the machine is considered to have higher quality than a doctor has. Conditional on a verdict that the machine wins, the second question pertains to which point on the machine ROC curve should be used for future decision making.

We tackle these two issues from both a Bayesian inference framework and a framework of frequentist classical analysis. In the Bayesian setting, we compute the posterior probability that the individual TPR/FPR pair lies below the ROC and search for a point on the ROC that dominates both the TPR and FPR with the largest posterior probability. Whether to declare the machine a winner is determined by a level of confidence that is chosen a priori. In the classical frequentist setting, we formulate the choice using both hypothesis testing and confidence region formation. The classical approach also requires a confidence level that is chosen a priori.

A uniform decision of relying exclusively only on machine or humans is likely to less optimal than an alternative framework that can differentiate the subset of humans who correctly process more information than that is available to the machine algorithm from those who are dominated by the machine algorithm. For example, Liang et al. (2019) proposes a machine learning model to make diagnosis on some childhood diseases. Comparison of the machine model with physicians (physicians are grouped according to their practice experience years) shows that their model outperforms the junior physician groups, but marginally underperforms senior physician groups. It is essential to design a rigorous framework to compare individual decision makers with machines. To the best of our knowledge, such a framework has yet to be developed. The goal of our paper is to fill this gap and formalize such a statistical approach. Our framework selects a winner between each human decision maker and the machine algorithm, and and applies the winning decision maker to classify future observations. In other words, future decision making is done through a combination of capable humans and the machine algorithm. This can be done through an AI assisted doctor's decision rule: the less-capable doctors (compared to AI algorithm) *always* follow the suggestion of artificial intelligence (or replaced by artificial intelligence), whereas skillful doctors (similar or better than AI algorithm) do not follow.

We apply our statistical framework to analyze doctor's medical diagnosis decision. In particular, we employ a unique dataset of National Free Pre-Pregnancy Checkups (NFPC) that dates back to 2010. This is a free health checkup project for conceiving couples and is conducted across 31 provinces in China. The data set includes doctor's ID and diagnosis of adverse pregnancy outcomes. We first split the overall dataset into two equal parts. The first part is used to compare doctors with machines. In particular, we employ a random

Electronic copy available at: https://ssrn.com/abstract=3508224

forest based method for diagnosis of risky pregnancy, which achieve an area under curve (AUC) of 0.6834. Under 95% confidence level, our statistical framework suggests that the random forest algorithm outperforms 44% doctors. We also find that doctors from rural area and under-developed cities have a relatively worse performance relative to others. We then replace less capable doctors with machines. The second dataset is used to validate a combined decision making strategy only by the more capable doctors and machine algorithms. In the second data set, the combined decision making achieves an increase of of 43.8% on TPR (the rate of correctly judging birth defect as high-risk pregnancy) and a reduction of 9.4% on FPR (the rate of misjudgment of normal birth as high-risk pregnancy) under Bayesian approach. The potential of improving accuracy using machine learning algorithms is substantial.

Our paper relates to at least two strands of literatures. First, it reflects on decision making with machine learning technology. For example, Wang et al. (2016) shows the usage of multitask machine learning algorithms in solving high-dimensional coordination decision-making policies for large-scale manufacturing systems. Berk (2017) analyzes the impact of machine learning tools on helping parole boards get better decision. Everett et al. (2018) creates decision-making agents for robotic motion planning using reinforcement learning methods. Kleinberg et al. (2018) documents that the usage of machine learning mechanisms can help judges in making good bail decisions. Nieto et al. (2019) finds that using machine learning algorithms can support university managers in academic decision making. Feng et al. (2019) researches the properties of ROC curves in decision making and presents a statistical inference framework for the ROC curve.

Second, our statistical framework of combining machines with humans in decision making relates to the economic analysis of the human and robot race. Since Keynes (2010), economists contemplates about the consequence of replacing human workers with new technology introduction. Recently, Acemoglu and Restrepo (2017) discover that rapidly aging countries are more likely to adopt industrial robots. Acemoglu and Restrepo (2018b) analyze the shape of the economy if new technologies such as AI will make some manual labor redundant. Acemoglu and Restrepo (2018a) investigate the differential impact on productivity and wage structure when machines take over the work of either low-skill labors or high-skill labors. Brynjolfsson et al. (2018) analyze the suitability for machine learning in different kinds of jobs and propose that achieving the potential of machines requires reforms of existing job content.

The rest of this paper are as follows. Section 2 presents the statistical model of human machine comparison. Section 3 provides the data and machine algorithm in this study. Section 4 conducts the empirical analysis of machine algorithms and human decision makers. Section 5 concludes.

Electronic copy available at: https://ssrn.com/abstract=3508224

2 Machine Learning and Human Decision Making

In a sample dataset with labels $Y_i \in \{0,1\}$, and associated features $X_i, i = 1, \ldots, n$:

$$\text{TPR} = \sum_{i=1}^{n} Y_i \widehat{Y}_i / \sum_{i=1}^{n} Y_i, \text{FPR} = \sum_{i=1}^{n} (1 - Y_i) \widehat{Y}_i / \sum_{i=1}^{n} (1 - Y_i).$$

where $\hat{Y}_i \in \{0,1\}$ is a predictor for Y_i. However, to account for sampling error, we need to introduce population analogs of TPR and FPR.

In order to make a comparison between decision maker and machines we need to make the following two assumptions:

Assumption 1. *The machine ROC represents a propensity score model of prediction:* $\mathbb{1}(p(x) \geqslant c_1)$. *The decision maker's decision is also based on a model* $\mathbb{1}(q(x) \geqslant c_2)$. $p(\cdot)$ *and* $q(\cdot)$ *can be correctly specified or misspecified. There is no incentive-features dependence across cases for each individual decision maker:* c_2 *does not depend on* x*, i.e. each decision maker uses the same weights for type I and II errors for all his/her cases.*

Assumption 2. *The machine ROC is precisely known without any sampling error.*

We have assumption 1 since difficulties with interpreting the ROC arise if individual decision makers (e.g. doctors) adopt difficult weights for trading off size and power, or when a given doctor adopt difficult weights for trading off size and power for different cases based on observed and unobserved features, as Appendix A.1 proposes. [2] Therefore, it is necessary to assume away incentive-feature dependence for individual decision maker before a meaningful comparison between machine and human can be made.

Assumption 2 can be justified by the fact that the sample size used to estimate the machine ROC is typically orders of magnitude larger than the number of cases for an individual decision maker. For instance, more than 1 million cases are used to estimate the ROC curve in our birth defect example, whereas the average cases per decision maker (doctor) is only about 40.

Sampling errors arise from the fewer number of observations for each individual decision maker. We focus on one individual decision maker, and we are interested in the parameter of the population pair of true positive and false positive rates for this decision maker:

[2] Related properties of machine ROC and human FPR/TPR pairs under incentive heterogeneity and information asymmetry are also discussed in Appendix. A.1

Electronic copy available at: https://ssrn.com/abstract=3508224

$\theta_0 = (\text{HFPR}_0, \text{HTPR}_0)$, where

$$\begin{aligned}\text{HTPR}_0 =& \frac{1}{p}\mathbb{E}Y_i\hat{Y}_i = \frac{1}{p}\int\int y\mathbb{1}\left(q\left(x\right) \geqslant c_2\right)\mathrm{d}y\mathrm{d}x,\\ \text{HFPR}_0 =& \frac{1}{1-p}\mathbb{E}\left(1-Y_i\right)\hat{Y}_i = \frac{1}{1-p}\int\int\left(1-y\right)\mathbb{1}\left(q\left(x\right) \geqslant c_2\right)f\left(y,x\right)\mathrm{d}y\mathrm{d}x.\end{aligned}$$

2.1 Comparison between Human FPR/TPR Pairs and Machine ROC Curves

Figure 2 shows the ROC curve of an AI algorithm and a FPR/TPR pair for a human decision maker. If the human FPR/TPR pair θ_0 is above the machine's ROC curve, it means that given HFPR_0, the machine has a lower TPR than human has; and given HTPR_0, machine has a larger FPR than human has. In this sense, human is better than the machine algorithm.[3]

[Figure 2 about here.]

If the human FPR/TPR point is below the machine ROC curve, one can always find a point A on the machine ROC curve matching human HFPR_0 but with a higher TPR. By the same token, one can find a point B on the machine ROC curve matching human HTPR_0 but with a lower FPR. Any point between point A and B on the machine ROC curve has a larger TPR and a smaller FPR than the human (HFPR_0,HTPR_0) pair; in this sense machine decision rule corresponding to a ROC curve from A to B is *better* than the human decision maker. The segment of the curve to the left of B has a smaller FPR but a lower TPR than human does. Points on this segment are not comparable to the human decision maker. By the same token, the segment of curve to the right of A has much higher TPR but larger FPR than human does, hence is neither superior nor inferior to the human. Lemma 2.1 provides a formal statement.

Lemma 2.1. *Denote ($HFPR_0$,$RTPR_0$) and ($RFPR_0$,$HTPR_0$) as two points on the machine's ROC curve sharing the same $HFPR_0$ and $HTPR_0$ of human decision maker respectively, if ($HFPR_0$,$HTPR_0$) is below the ROC curve, then the portion of the ROC curve between ($HFPR_0$,$RTPR_0$) and ($RFPR_0$,$HTPR_0$) corresponds to better decision rules.*

Suppose now θ_0 lies below the ROC, to be more precise about the sense by which the machine is "better" than the decision maker, recall that θ_0 is a population parameter, measuring the average human performance after seeing an infinite number of cases. This

[3] With assumption 1, θ_0 lies above the ROC if human correctly uses at least as much information as the machine.

Electronic copy available at: https://ssrn.com/abstract=3508224

only means that *on average*, the machine can improve on the decision maker by increasing TPR given FPR, or reducing FPR given TPR. However, unless the machine can make a judgment with perfect accuracy, for a given case, the decision maker can either underperform or outperform the machine. From this point on, we will agree that our goal is to find out whether θ_0 is above or below the ROC, or whether the machine is better than the individual decision maker *on average*.

Typically we do not observe θ_0, even though we can estimate it using a sample of observations:

$$\hat{\theta} = (\text{FPR}, \text{TPR})$$

where, with $\hat{Y}_i = \mathbb{1}\left(q\left(X_i\right) \geqslant c_2\right)$,

$$\text{TPR} = \sum_{i=1}^{n} Y_i \hat{Y}_i / \sum_{i=1}^{n} Y_i, \quad \text{FPR} = \sum_{i=1}^{n} \left(1 - Y_i\right) \hat{Y}_i / \sum_{i=1}^{n} \left(1 - Y_i\right).$$

In the presence of sampling uncertainty, the inference problem pertains to how do we make probabilistic statement regarding whether θ_0 is above or below the ROC. Note that $\hat{\theta}$ is a vector function of a multinomial distribution. It can be simulated but does not admit an analytical density function or distribution function. Large sample analysis however, is facilitated by the joint asymptotic distribution of the FPR/TPR pair.

Lemma 2.2. *The joint asymptotic distribution of a human's FPR and TPR pair is multivariate normal.*

The proof of lemma 2.2 is given in Appendix A.3. Essentially it consists of an application of the Delta method to a bivariate function of a four-dimensional multinomial distribution, which itself is asymptotically multivariate normal.

A word of caution about the asymptotic distribution of

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \Omega\right).$$

Note that this is under the assumption that human behavioral rule is

$$\hat{Y}_i = \mathbb{1}\left(q\left(X_i\right) \geqslant c_2\right)$$

In other words, decision makers know their models already before applying to the ob-

Electronic copy available at: https://ssrn.com/abstract=3508224

servations in the dataset. [4] As a consequence of Lemma 2.2, a typical confidence set of FPR/TPR pair with α confidence level is asymptotically of oval-shape. As shown in Figure 3, the blue ellipse is an 95% confidence level for a decision maker in our data set described in the next section. The yellow points are the bootstrapped FPR/TPR pairs.

[Figure 3 about here.]

We provide two approaches to address the question of sampling uncertainty. The first is based on Bayesian analysis principle, and the second is based on classical frequentist inference.

2.2 The Bayesian approach

The Bayesian approach is potentially easier to interpret than the classical frequentist method is. A Bayesian method requires specifying a prior distribution for the decision maker's θ, and the likelihood of the data given θ. Recall that this model is completely specified by three parameters: $p_1 = \mathbb{E}Y\hat{Y}$, $p_2 = \mathbb{E}\left(1-Y\right)\hat{Y}$, and $p_3 = \mathbb{E}Y\left(1-\hat{Y}\right)$. The data follows a multinomial distribution with four categories. Conceptually, $\theta = h\left(p\right)$, $h = \left(h_1, h_2\right)$. $\hat{p} = \left(\hat{p}_1, \hat{p}_2, \hat{p}_3\right)$ are sufficient statistics summarizing the data. The multinomial distribution is a completely specified parametric model. It allows for exact likelihood based Bayesian posterior distribution computation. Given p, the data likelihood is

$$L\left(Y|p\right) = p_1^{n\hat{p}_1} p_2^{n\hat{p}_2} p_3^{n\hat{p}_3} \left(1-p_1-p_2-p_3\right)^{n\left(1-\hat{p}_1-\hat{p}_2-\hat{p}_3\right)}.$$

Given a prior $\pi\left(p\right)$, the posterior distribution can be computed (e.g. by simulation)

$$p\left(p|Y\right) \propto \pi\left(p\right) L\left(Y|p\right).$$

Projecting the simulated $p\left(p|Y\right)$ onto $\theta = h\left(p\right)$ gives a simulated posterior distribution for θ. Credible regions can be constructed by smoothing or contouring.

It is well known that Dirichlet prior is a conjugate prior for multinomial distribution to allow for analytic posterior calculation. It includes as a special case the uniform distribution over p, and also includes the Beta distribution as a special case when p is one dimensional. To our knowledge using Dirichlet prior to conduct posterior Bayesian inference in ROC curve is a new contribution.

Consider a Dirichlet prior on the $K(=4)$ simplex $p_1, \ldots, p_K$, such that $p_1 + \ldots + p_K = 1$, with parameters $\alpha_1, \ldots, \alpha_K$. $\alpha_1 = \ldots = \alpha_K = \alpha$ is called a symmetric Dirichlet

[4] There is no model estimation uncertainty for the decision makers, and the decision makers do not try to learn a better model to improve prediction from the sample observations. In the presence of decision makers learning from the data, the analysis will be much more complex.

Electronic copy available at: https://ssrn.com/abstract=3508224

distribution. Given symmetry, $\alpha = 1$ corresponds to a uniform prior on $p_1, \ldots, p_K$. The posterior distribution is also Dirichlet with parameters

$$\hat{\alpha}_k = \alpha_k + n\hat{p}_k, \quad k = 1, \ldots, K.$$

with uniform prior, the posterior Dirichlet distribution has parameters [5]

$$\hat{\alpha}_k = 1 + n\hat{p}_k, \quad k = 1, \ldots, K.$$

To simulate the posterior distribution of $\theta = (\theta_1, \theta_2)$, for $s = 1, \ldots, S$, make draws from the Dirichlet distribution with parameters $\hat{\alpha}_k = 1 + n\hat{p}_k, \quad k = 1, \ldots, K$. This is generated by setting

$$p_{sk} = y_{sk} / \sum_{j=1}^{K} y_{js}, \quad k = 1, \ldots, K,$$

such that $y_{sk} = \sum_{t=1}^{\hat{\alpha}_k} \mathcal{E}_{ts}$ and $\mathcal{E}_{ts}$ are unit exponential variables independent from each other. Conventional confidence and credibility levels suggest that we are conservative when making decisions to replace humans with machines. Specifically, only w For each $p_s = (p_{sk}, k = 1, \ldots, K)$, calculate

$$\theta_s = h\left(p_s\right).$$

We can calculate the posterior probability that the θ lies below the ROC:

$$\int_{\theta \text{ below ROC}} p\left(\theta|Y\right) d\theta.$$

or the posterior probability of human decision maker being dominated by the ROC, by the fraction of times where θ_s lies below the ROC.

In addition to the posterior probability that θ lies below the ROC, we are also interested in the *maximum* posterior probability of a set of θ that are dominated simultaneously by a single point on the ROC:

$$\text{maxposteriorprob} \equiv \sup_{a \in (0,1)} \int_{\theta_1 \geqslant a, \theta_2 \leqslant g(a)} p\left(\theta|Y\right) d\theta. \tag{1}$$

where the machine ROC has been represented by $\theta_2 = g\left(\theta_1\right)$. Both of these can be computed by simulation: for $\theta_r, r = 1, \ldots, R$ simulated draws from the posterior distribution,

[5] For more information see https://en.wikipedia.org/wiki/Dirichlet_distribution

Electronic copy available at: https://ssrn.com/abstract=3508224

they are estimated by

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{1}\left(\theta_r \text{ below ROC}\right)$$

and

$$\sup_{a\in(0,1)}\frac{1}{R}\sum_{r=1}^{R}\mathbb{1}\left(\theta_{1r}\geqslant a,\ \theta_{2r}\leqslant g\left(a\right)\right).$$

An alternative to exact finite sample Bayesian analysis is a semiparametric asymptotic Bayesian method in Kim (2002), which combines a prior distribution for θ with the approximate normal distribution of $\hat{\theta}$ around θ to form a posterior distribution. This method can be formally justified by the convergence of the finite sample Bayesian posterior distribution to the approximate normal limit, Chernozhukov and Hong (2003). Let $\pi\left(\theta\right)$ be a prior distribution for θ, e.g. the uniform distribution on the two dimensional unit square. It has been shown that

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}N\left(0,\Omega\right)\quad\text{or}\quad\hat{\theta}\overset{A}{\approx}N\left(\theta,\frac{\hat{\Omega}}{n}\right).$$

where Ω can be estimated analytically or by the bootstrap. The entire distribution of $f_{\hat{\theta}-\theta_0}\left(\cdot\right)$ can also be estimated by the bootstrap. The above convergence in distribution can be strengthened to convergence in the moment total variation norm in Chernozhukov and Hong (2003). The asymptotic posterior distribution of θ is defined through

$$\begin{aligned}p\left(\theta|\hat{\theta}\right)\propto&\pi\left(\theta\right)\hat{f}_{\hat{\theta}-\theta_0}\left(\hat{\theta}-\theta\right)\\\approx&\pi\left(\theta\right)\exp\left(-\frac{n}{2}\left(\theta-\hat{\theta}\right)'\hat{\Omega}^{-1}\left(\theta-\hat{\theta}\right)\right).\end{aligned}$$

When a (diffuse) uniform prior is chosen, $\pi\left(\theta\right)\equiv 1$,

$$\begin{aligned}p\left(\theta|\hat{\theta}\right)\propto&\hat{f}_{\hat{\theta}-\theta_0}\left(\hat{\theta}-\theta\right)\\\approx&\exp\left(-\frac{n}{2}\left(\theta-\hat{\theta}\right)'\hat{\Omega}^{-1}\left(\theta-\hat{\theta}\right)\right).\end{aligned}$$

The $1-\alpha$ confidence region $\hat{S}$ constructed either by the approximate normal distribution or by bootstrapping therefore corresponds to the $1-\alpha$ Bayesian credible region using a uniform prior distribution and similar constructions. A variety of Bayesian techniques (e.g. highest density regions) can be used to constructed posterior credible regions using possibly non-

Electronic copy available at: https://ssrn.com/abstract=3508224

uniform prior densities. These often involve the use of simulation methods, such as MCMC. The normal approximation does present the drawback that the support is not limited to $(0,1)$. However for large sample sizes confidence sets at conventional levels should well reside within $(0,1)$ with large probabilities.

2.3 The Frequentist approach

Consider the following heuristic procedure.

- Form a confidence set $\hat{S}$ for θ_0, for example based on bootstrapping or the asymptotic normal distribution of $\hat{\theta}$ around θ_0.
- For each $s \in \hat{S}$, define $A_s = \{\theta : \theta_1 \leqslant s_1, \theta_2 \geqslant s_2\}$. In other words, A_s is the set of points that are *better* than s (namely, lower FPR and higher TPR).
- Next define

$$A = \Cap_{s\in\hat{S}} A_s$$

  A is the set of points that are *better* than all the points in $\hat{S}$ in the above sense.

- Next define

$$\bar{A} = A \cap \text{ROC}$$

  Namely, $\bar{A}$ is the set of points on the ROC that simultaneously dominates all the points in $\hat{S}$.

How do we provide a statistical probabilistic statement justifying this procedure? In a classical approach, θ_0 is a fixed number and not a random variable. It is either above or below the ROC, without an associated probability. The confidence set itself is a random set, such that

$$P\left(\theta_0 \in \hat{S}\right) \approx 1-\alpha.$$

Namely, if the world is to repeat 100 times, about 95 times $\theta_0 \in \hat{S}$. Consequently, $\bar{A}$ is a random set, that is possibly empty, such that

$$P\left(\text{all points on } \bar{A} \text{ dominates } \theta_0\right) \approx 1-\alpha.$$

Electronic copy available at: https://ssrn.com/abstract=3508224

What about the probability of the random event of $A \cap \text{ROC} \neq \varnothing$? Obviously

$$P_{\theta_0}\left(A \cap \text{ROC} \neq \varnothing\right)$$

depends on θ_0 and is unknown. It can be estimated by replacing θ_0 with $\hat{\theta}$ and iterates on the bootstrap procedure. Depending on where the true θ_0 lies, $P_{\theta_0}\left(A \cap \text{ROC} \neq \varnothing\right)$ can be anywhere between 0 and 1.

In the Bayesian setting, the event of $A \cap \text{ROC} \neq \varnothing$ then has the interpretation that $1-\alpha$ portion of the posterior distribution of θ lies below the ROC. For example, there exists a point on the ROC that simultaneously dominates 95% of the posterior distribution of θ. A Bayesian method can say more than constructing a posterior credible region. As discussed before, the posterior probability that θ lies below the ROC can be computed by numerical integration or estimated by simulations.

In the classical/frequentist framework, θ_0 is a fixed number and not a random variable, and we are not allowed to make probabilistic statement about it. We therefore turn to a Hypothesis testing setup, to formulate tests of

$$H_0 : \theta_0 \text{ lies below ROC} \quad \text{against} \quad H_1 : \theta_0 \text{ lies above ROC}.$$

The roles of the null and alternative hypotheses can be reversed, depending on which one we are more willing to theorize on, *a priori*. This is a test of composite null versus composite alternative hypotheses. In general, finding the uniformly most powerful test is difficult.

Suppose we represent the ROC curve by $\theta_2 = g\left(\theta_1\right)$, where $g\left(\cdot\right)$ is known, increasing, and concave. Then we can rewrite

$$H_0 : \theta_{20} \leqslant g\left(\theta_{10}\right) \quad \text{against} \quad H_1 : \theta_{20} \geqslant g\left(\theta_{10}\right).$$

Let $t = \theta_2 - g\left(\theta_1\right)$, then the hypotheses can be restated as

$$H_0 : t_0 \leqslant 0 \quad \text{against} \quad H_1 : t_0 \geqslant 0.$$

An asymptotic test can then be based on combining the Delta method with a one-sided t-test.

For $G_0 = g'\left(\theta_{10}\right)$, $B = \begin{pmatrix} 1 & G_0 \end{pmatrix}$, it can be shown that

$$\sqrt{n}\left(\hat{t} - t_0\right) \xrightarrow{d} N\left(0, B\Omega B'\right).$$

Electronic copy available at: https://ssrn.com/abstract=3508224

Let $\hat{G} = g'\left(\hat{\theta}_1\right)$, $\hat{B} = \begin{pmatrix} 1 & \hat{G} \end{pmatrix}$, $\hat{\Omega} \xrightarrow{p} \Omega$, and $z_{1-\alpha}$ the standard normal $1-\alpha$ percentile:

$$P\left(\frac{\sqrt{n}\left(\hat{t} - t_0\right)}{\sqrt{\hat{B}\hat{\Omega}\hat{B}'}} \geqslant z_{1-\alpha}\right) \to \alpha$$

An approximate size α test then implies rejection when $\sqrt{n}\hat{t}/\sqrt{\hat{B}\hat{\Omega}\hat{B}'} > z_{1-\alpha}$. For conventional levels of α, typically we do not reject when $\hat{t} \leqslant 0$.

A typically test rejects when $\hat{t} \geqslant \hat{c}_{1-\alpha}$, for a suitable chosen $\hat{c}_{1-\alpha}$ such that

$$P\left(\hat{t} \geqslant \hat{c}_{1-\alpha} | t_0 = 0\right) \to \alpha.$$

The Delta method sets $\hat{c}_{1-\alpha} = \sqrt{\frac{\hat{B}\hat{\Omega}\hat{B}'}{n}} z_{1-\alpha}$ There are alternative ways to implement $\hat{c}_{1-\alpha}$. One is the so-called asymptotic delta method: for large R, draw for $r = 1, \ldots, R$, ω_r from $N\left(0, \frac{\hat{\Omega}}{n}\right)$. Set $\hat{c}_{1-\alpha}$ as the $1-\alpha$ percentile of the empirical distribution of

$$\hat{\theta}_2 + \omega_{2r} - g\left(\hat{\theta}_1 + \omega_{1r}\right) - \left(\hat{\theta}_2 - g\left(\hat{\theta}_1\right)\right)$$

Another alternative is bootstrapping. Set $\hat{c}_{1-\alpha}$ to the $1-\alpha$ percentile of the empirical distribution of $\hat{t}_b - \hat{t}, b = 1, \ldots, B$, where $\hat{t}_b$'s are bootstrapped replications of $\hat{t}$.

For completeness, we also describe the reverse testing problem:

$$H_0 : \theta_0 \text{ lies above ROC} \quad \text{against} \quad H_1 : \theta_0 \text{ lies below ROC.}$$

This is appropriate when we only replace a human decision maker with the machine when there is *overwhelming* evidence that the machine is better. A priori, we hold confidence in human being, unless there is strong evidence suggesting otherwise.

$$H_0 : \theta_{20} \geqslant g\left(\theta_{10}\right) \quad \text{against} \quad H_1 : \theta_{20} \leqslant g\left(\theta_{10}\right).$$

Let $t = \theta_2 - g\left(\theta_1\right)$, then the hypotheses can be restated as

$$H_0 : t_0 \geqslant 0 \quad \text{against} \quad H_1 : t_0 \leqslant 0.$$

We reject if $\hat{t} \leqslant \hat{c}_\alpha$. There are several choices for $\hat{c}_\alpha$.

- Choice 1:

$$\hat{c}_\alpha = z_\alpha \sqrt{\frac{\hat{B}\hat{\Omega}\hat{B}'}{n}}.$$

Electronic copy available at: https://ssrn.com/abstract=3508224

- Choice 2: asymptotic delta method: for large R, draw for $r = 1, \ldots, R$, ω_r from $N\left(0, \frac{\hat{\Omega}}{n}\right)$. Set $\hat{c}_\alpha$ as the α percentile of the empirical distribution of

$$\hat{\theta}_2 + \omega_{2r} - g\left(\hat{\theta}_1 + \omega_{1r}\right) - \left(\hat{\theta}_2 - g\left(\hat{\theta}_1\right)\right)$$

- Bootstrapping. Set $\hat{c}_\alpha$ to the α percentile of the empirical distribution of $\hat{t}_b - \hat{t}, b = 1, \ldots, B$, where $\hat{t}_b$'s are bootstrapped replications of $\hat{t}$.

The next question regards whether the constructed confidence set can be used to test H_0 against H_1. Recall that a conventional confidence set is given asymptotically by

$$\{\theta : n\left(\theta - \hat{\theta}\right)' \Omega^{-1} \left(\theta - \hat{\theta}\right) \leqslant \chi^2_{2,1-\alpha}\},$$

Now consider two tests.

- Test 1 T1: reject if $\hat{\theta}_2 > g\left(\hat{\theta}_1\right)$ and $\left(\hat{\theta}_1, g\left(\hat{\theta}_1\right)\right)$ is not in the confidence set.
- Test 2 T2: reject if $\hat{\theta}_2 > g\left(\hat{\theta}_1\right)$ and no point of the ROC is in the confidence set.

T1 rejects if $\hat{\theta}_2 > g\left(\hat{\theta}_1\right)$ and if

$$n\left[\hat{\Omega}^{-1}\right]_{2,2} \left(\hat{\theta}_2 - g\left(\hat{\theta}_1\right)\right)^2 \geqslant \chi^2_{2,1-\alpha},$$

This test is unlikely to have the correct size. By partitioned matrix inversion

$$\left[\hat{\Omega}^{-1}\right]^{-1}_{2,2} = \sigma_{22} - \frac{\sigma^2_{12}}{\sigma_{11}}$$

However, under the null:

$$\hat{\theta}_2 - g\left(\hat{\theta}_1\right) \approx \hat{\theta}_2 - \theta_2 - G_0\left(\hat{\theta}_1 - \theta_1\right) \sim N\left(0, \sigma_{22} + G_0^2\sigma_{11} - 2G_0\sigma_{12}\right).$$

The test statistic distribution is not χ^2_2. Of course size calibration is possible, but then it amounts to the asymptotic one-sided t-test.

T2 rejects if $\hat{\theta}_2 > g\left(\hat{\theta}_1\right)$ and if

$$\inf_{x: x_2 = g(x_1)} n\left(x - \hat{\theta}\right)' \hat{\Omega} \left(x - \hat{\theta}\right) \geqslant \chi^2_{2,1-\alpha}.$$

The left hand side is the J-statistic for one overidentifying restriction, and has an asymptotic χ^2_1 distribution (one degree of freedom, not two). Since $\chi^2_{2,1-\alpha} \geqslant \chi^2_{1,1-\alpha}$, the null

Electronic copy available at: https://ssrn.com/abstract=3508224

rejection probability is less than α. Therefore this test is conservatively valid (undersized), and is likely to be less powerful. This is in addition to the power loss from essentially using a two sided test for a one-sided hypothesis. It is again possible to calibrate size.

Next, test T3 rejects if $\hat{\theta}_2 > g\left(\hat{\theta}_1\right)$ and if there exists a point on the ROC that has worse TPR and FPR simultaneously for all points in the confidence set. T3 is obviously even more conservative than T2. Therefore it will be undersized, and less powerful. These theoretical properties can all be validated in Monte Carlo simulations.

There are many alternative ways to construct confidence sets $\hat{S}$ such that

$$P\left(\theta_0 \in \hat{S}\right) \approx 1-\alpha.$$

Each corresponds to a construction of $\bar{A} = A \cap \text{ROC}$. Which one leads to a "better" $\bar{A}$ holding α constant?

Perhaps the length of $\bar{A}$ can be optimized. Since it is random, perhaps use the expected length of $\bar{A}$. Denote the ROC by $\theta_2 = g(\theta_1)$. Suppose $\bar{A}$ is formed by two end points $\{g^{-1}(b), b\}$ and $\{a, g(a)\}$. Then

$$\text{len}\left(\bar{A}\right) = \int_{g^{-1}(b)}^{a} \sqrt{(dx)^2 + (g'(x))\, dx^2} = \int_{g^{-1}(b)}^{a} \sqrt{1 + g'(x)^2}dx.$$

Now consider confidence sets of the form of

$$\hat{S} = \{\theta : \theta_1 \geqslant \hat{\theta}_1 - a, \theta_2 \leqslant \hat{\theta}_2 + b\}$$

This implies that we choose a and b such that

$$P\left(\hat{\theta}_1 - \theta_{10} \leqslant a, \hat{\theta}_2 - \theta_{20} \geqslant -b\right) \to 1-\alpha$$

The joint distribution of $\hat{\theta} - \theta_0$ can be simulated to find a and b for a given α. To determine a and b, one possibility is to limit to $a = b$:

$$P\left(\hat{\theta}_1 - \theta_{10} \leqslant a, \hat{\theta}_2 - \theta_{20} \geqslant -a\right) \to 1-\alpha$$

Another possible is to maximize expected length of $\bar{A}$:

$$\begin{aligned} &\max_{a,b} \mathbb{E} \int_{g^{-1}\left(\hat{\theta}_2+b\right)}^{\hat{\theta}_1-a} \sqrt{1+g'(x)^2}dx \\ &\text{such that} \quad P\left(\hat{\theta}_1 - \theta_{10} \leqslant a, \hat{\theta}_2 - \theta_{20} \geqslant -a\right) \approx 1-\alpha \end{aligned}$$

Electronic copy available at: https://ssrn.com/abstract=3508224

This program depends on the true underlying data generating process, which is unknown. A feasible version can be estimated by bootstrapping: Let $\mathbb{E}^*$, P^*, $\hat{\theta}^*$ denote bootstrap versions that can be simulated:

$$\max_{a,b} \mathbb{E}^* \int_{g^{-1}\left(\hat{\theta}_2^*+b\right)}^{\hat{\theta}_1^*-a} \sqrt{1+g'(x)^2}dx$$
$$\text{such that} \quad P^*\left(\hat{\theta}_1^* - \hat{\theta}_1 \leqslant a, \hat{\theta}_2^* - \hat{\theta}_2 \geqslant -b\right) = 1-\alpha$$

This is likely to generate a longer $\bar{A}$ given a fixed confidence level.

More generally, we can construct "two dimensional one-sided" confidence sets by inverting two-dimensional inequality tests of the form of

$$H_0 : \theta_1 \leqslant \theta_{10} \quad \text{and} \quad \theta_2 \geqslant \theta_{20} \quad \text{against} \quad H_1 : \text{not } H_0.$$

For example a test may reject when

$$\hat{\theta}_1 \geqslant \theta_{10} + a \quad \text{or} \quad \hat{\theta}_2 \leqslant \theta_{20} - b.$$

where a and b are obtained from simulating the limiting distribution. This corresponds then to the confidence set of the form of

$$\hat{S} = \{\theta : \theta_1 \geqslant \hat{\theta}_1 - a \quad \text{and} \quad \theta_2 \leqslant \hat{\theta}_2 + b\}.$$

Reverse versions of the conservative tests T2 and T3 take the following form. T2 rejects if $\hat{\theta}_2 < g\left(\hat{\theta}_1\right)$ and if

$$\inf_{x:x_2=g(x_1)} n\left(x-\hat{\theta}\right)' \hat{\Omega}\left(x-\hat{\theta}\right) \geqslant \chi^2_{2,1-\alpha}.$$

T3 rejects if $\hat{\theta}_2 < g\left(\hat{\theta}_1\right)$ and if there exists a point on the ROC that has better TPR and FPR simultaneously for all points in the confidence set.

Recall that the Bayesian method is able to find the *best* point on the ROC. It is harder for the frequentist approach to do so, since typically in a frequentist approach, the level of confidence is held fixed by convention. In the classical setting, there can be several heuristic procedures for this purpose, supposing that $\hat{\theta}$ is below the ROC.

In the first approach, we look for pairs (a, b), such that (a, b) lies on the ROC, that the

Electronic copy available at: https://ssrn.com/abstract=3508224

following level of confidence in the superiority of the machine algorithm:

$$P\left(\hat{\theta}_1 \geqslant a, \hat{\theta}_2 \leqslant b\right)$$

This coverage probability can be estimated by bootstrapping

$$P^*\left(\hat{\theta}_1^* \geqslant a, \hat{\theta}_2^* \leqslant b\right). \tag{2}$$

An implementable program is therefore

$$\max_{b=g(a)} P^*\left(\hat{\theta}_1^* \geqslant a, \hat{\theta}_2^* \leqslant b\right). \tag{3}$$

Conventional frequentist confidence set formation centers around θ_0. Therefore we consider two alternative approaches. In the second approach, we want to maximize over pairs (a, b), such that $\left(\hat{\theta}_1 - a, \hat{\theta}_2 + b\right)$ lies on the ROC, the level of confidence in the superiority of the machine algorithm over the human decision maker:

$$P\left(\hat{\theta}_1 - a \leqslant \theta_{10},\ \hat{\theta}_2 + b \geqslant \theta_{20}\right) = P\left(\hat{\theta}_1 - \theta_{10} \leqslant a,\ \hat{\theta}_2 - \theta_{20} \geqslant -b\right)$$

This coverage probability can be estimated by bootstrapping

$$P^*\left(\hat{\theta}_1^* - \hat{\theta}_1 \leqslant a,\ \hat{\theta}_2^* - \hat{\theta}_2 \geqslant -b\right)$$

for each $a \in \left(0, \hat{\theta}_1 - g^{-1}\left(\hat{\theta}_2\right)\right)$, $b = g\left(\hat{\theta}_1 - a\right) - \hat{\theta}_2$. An implementable program is

$$\max_{a \in \left(0, \hat{\theta}_1 - g^{-1}\left(\hat{\theta}_2\right)\right)} P^*\left(\hat{\theta}_1^* - \hat{\theta}_1 \leqslant a,\ \hat{\theta}_2^* - \hat{\theta}_2 \geqslant -\left(g\left(\hat{\theta}_1 - a\right) - \hat{\theta}_2\right)\right).$$

This can also be written as

$$\max_{a \in \left(\hat{\theta}_1, 2\hat{\theta}_1 - g^{-1}\left(\hat{\theta}_2\right)\right)} P^*\left(\hat{\theta}_1^* \leqslant a,\ \hat{\theta}_2^* \geqslant 2\hat{\theta}_2 - g\left(2\hat{\theta}_1 - a\right)\right). \tag{4}$$

If the maximized value is larger than the given confidence level, then the maximizing pair of (a, b) on the ROC is chosen. Otherwise do not replace with the machine result.

This procedure can be given the following interpretation: $\left(\hat{a}, \hat{b}\right)$ is a consistent estimate of (a_0, b_0), where (a_0, b_0) is a point on the ROC where the probability of

$$P\left(\hat{\theta}_1 \leqslant \theta_{10} + a, \hat{\theta}_2 \geqslant \theta_{20} - b\right)$$

is maximized with respect to a and b. Treat ROC and (the unknown) θ_0 as fixed, and θ_0

Electronic copy available at: https://ssrn.com/abstract=3508224

can be on either sides of the ROC.

$$\text{argmax}_{a\in(0,\theta_{10}-g^{-1}(\theta_{20})),b=g(\theta_{10}-a)-\theta_{20}} P\left(\hat{\theta}_1 \leqslant \theta_{10}+a, \hat{\theta}_2 \geqslant \theta_{20}-b\right)$$

The bootstrap analog provides an estimate $\left(\hat{a},\hat{b}\right)$ of (a_0,b_0) that are guaranteed to lie on the ROC.

Alternatively, we might want to minimize over these pairs (a,b) such that $\left(\hat{\theta}_1-a,\hat{\theta}_2+b\right)$ lies on the ROC the level of (lack of) confidence in the superiority of the machine algorithm:

$$P\left(\hat{\theta}_1-a\geqslant\theta_{10},\ \hat{\theta}_2+b\leqslant\theta_{20}\right)=P\left(\hat{\theta}_1-\theta_{10}\geqslant a,\ \hat{\theta}_2-\theta_{20}\leqslant -b\right)$$

The true DGP is unknown, but we can estimate the above coverage probability by bootstrapping

$$P^*\left(\hat{\theta}_1^*-\hat{\theta}_1\geqslant a,\ \hat{\theta}_2^*-\hat{\theta}_2\leqslant -b\right)$$

for each $a\in\left(0,\hat{\theta}_1-g^{-1}\left(\hat{\theta}_2\right)\right)$, $b=g\left(\hat{\theta}_1-a\right)-\hat{\theta}_2$. Thus an implementable program is

$$\min_{a\in\left(0,\hat{\theta}_1-g^{-1}\left(\hat{\theta}_2\right)\right)} P^*\left(\hat{\theta}_1^*-\hat{\theta}_1\geqslant a,\ \hat{\theta}_2^*-\hat{\theta}_2\leqslant -\left(g\left(\hat{\theta}_1-a\right)-\hat{\theta}_2\right)\right). \tag{5}$$

This can also be written as

$$\min_{a\in\left(\hat{\theta}_1,2\hat{\theta}_1-g^{-1}\left(\hat{\theta}_2\right)\right)} P^*\left(\hat{\theta}_1^*\geqslant a,\ \hat{\theta}_2^*\leqslant 2\hat{\theta}_2-\left(g\left(2\hat{\theta}_1-a\right)\right)\right). \tag{6}$$

If the minimized value is smaller than the one minus the given confidence level, then the minimizing pair of (a,b) on the ROC is chosen. Otherwise do not replace with the machine result.

This procedure can be given the following interpretation: $\left(\hat{a},\hat{b}\right)$ is a consistent estimate of (a_0,b_0), where (a_0,b_0) is a point on the ROC where the probability of

$$P\left(\hat{\theta}_1\geqslant\theta_{10}+a,\hat{\theta}_2\leqslant\theta_{20}-b\right)$$

is minimized with respect to a and b. Treat ROC and (the unknown) θ_0 as fixed, and θ_0 can be on either sides of the ROC.

$$\text{argmin}_{a\in(0,\theta_{10}-g^{-1}(\theta_{20})),b=g(\theta_{10}-a)-\theta_{20}} P\left(\hat{\theta}_1\geqslant\theta_{10}+a,\hat{\theta}_2\leqslant\theta_{20}-b\right)$$

The bootstrap analog provides an estimate $\left(\hat{a},\hat{b}\right)$ of (a_0,b_0) that are guaranteed to lie on

Electronic copy available at: https://ssrn.com/abstract=3508224

the ROC.

Test inversion and bootstrapping P-value Another approach is to consider the inversion of test statistics. First we formulate a null hypothesis for the superiority (or lack of it) of machine over humans, and a corresponding testing procedure. Next we use bootstrap to compute the P-value of the test, and locate the pair of $a, b = g(a)$ that optimizes the bootstrapped P-value in favor of a decision of replacing humans with machines. Example details are given below for four different setups.

Framework 1: For each $a, b = g(a)$. Let H_0 be $\{\theta_{10} \leqslant a \text{ or } \theta_{20} \geqslant b\}$. H_1 negates H_0. In other words, H_0 states that human is not worse than the machine. We want to minimize a bootstrapped P-value of a test procedure, since rejection means machine is better, and a small P-value means the data provides the strongest evidence for rejection.

A natural (although its optimality remains to be investigated) rejection area is

$$R = \{\hat{\theta}_1 \geqslant c_a \text{ and } \hat{\theta}_2 \leqslant c_b\}.$$

For each pair of c_a, c_b, the rejection probability under $\theta_0 = (a, b)$ (not proven yet to be the least favorable null),

$$P\left(\hat{\theta}_1 - a \geqslant c_a - a \text{ and } \hat{\theta}_2 - b \leqslant c_b - b\right)$$

can be estimated by bootstrap:

$$P^*\left(\hat{\theta}_1^* - \hat{\theta}_1 \geqslant c_a - a \text{ and } \hat{\theta}_2^* - \hat{\theta}_2 \leqslant c_b - b\right)$$

The bootstrapped P-value is thus

$$P^*\left(\hat{\theta}_1^* - \hat{\theta}_1 \geqslant \hat{\theta}_1 - a \text{ and } \hat{\theta}_2^* - \hat{\theta}_2 \leqslant \hat{\theta}_2 - b\right) = P^*\left(\hat{\theta}_1^* \geqslant 2\hat{\theta}_1 - a \text{ and } \hat{\theta}_2^* \leqslant 2\hat{\theta}_2 - b\right)$$

The program to implement is then

$$\min_{b=g(a)} P^*\left(\hat{\theta}_1^* \geqslant 2\hat{\theta}_1 - a \text{ and } \hat{\theta}_2^* \leqslant 2\hat{\theta}_2 - b\right) \tag{7}$$

The machine is chosen if the minimized value is smaller than the given size (e.g. 5%). Indeed, (7) is exactly the same program as (5) and (6).

Framework 2: For each $a, b = g(a)$. Let H_0 be $\{\theta_{10} \geqslant a \text{ or } \theta_{20} \leqslant b\}$. H_1 negates H_0. In other words, H_0 states that machine is not worse than human. Rejection means machine is worse than human, or human is better than machine. Do not replace human with machine in case of rejection, only replace in case of failing to reject. We want to maximize

Electronic copy available at: https://ssrn.com/abstract=3508224

a bootstrapped P-value of a test procedure, since rejection means human is better, and a large P-value means the data provides the weakest evidence for rejection, or the strongest evidence for replacing human with machine.

A natural (although its optimality remains to be investigated) rejection area is

$$R = \{\hat{\theta}_1 \leqslant c_a \text{ and } \hat{\theta}_2 \geqslant c_b\}.$$

For each pair of c_a, c_b, the rejection probability under $\theta_0 = (a, b)$ (not proven yet to be the least favorable null),

$$P\left(\hat{\theta}_1 - a \leqslant c_a - a \text{ and } \hat{\theta}_2 - b \geqslant c_b - b\right)$$

can be estimated by bootstrap:

$$P^*\left(\hat{\theta}_1^* - \hat{\theta}_1 \leqslant c_a - a \text{ and } \hat{\theta}_2^* - \hat{\theta}_2 \geqslant c_b - b\right)$$

The bootstrapped P-value is thus

$$P^*\left(\hat{\theta}_1^* - \hat{\theta}_1 \leqslant \hat{\theta}_1 - a \text{ and } \hat{\theta}_2^* - \hat{\theta}_2 \geqslant \hat{\theta}_2 - b\right) = P^*\left(\hat{\theta}_1^* \leqslant 2\hat{\theta}_1 - a \text{ and } \hat{\theta}_2^* - \geqslant 2\hat{\theta}_2 - b\right)$$

The program to implement is then

$$\max_{b=g(a)} P^*\left(\hat{\theta}_1^* \leqslant 2\hat{\theta}_1 - a \text{ and } \hat{\theta}_2^* \geqslant 2\hat{\theta}_2 - b\right) \tag{8}$$

The machine is chosen if the maximized value is greater than the given desired level. Indeed, (8) is exactly the same program as (4).

Framework 3: For each $a, b = g(a)$. Let H_0 be $\{\theta_{10} \leqslant a \text{ and } \theta_{20} \geqslant b\}$. H_1 negates H_0. In other words, H_0 states that human is better than the machine. Rejection means human is not better than machine. We want to minimize a bootstrapped P-value of a test procedure, since rejection means human is no better, and a small P-value means the data provides the strongest evidence for rejection.

A natural (although its optimality remains to be investigated) rejection area is

$$R = \{\hat{\theta}_1 \geqslant c_a \text{ or } \hat{\theta}_2 \leqslant c_b\}.$$

For each pair (c_a, c_b), the rejection probability under $\theta_0 = (a, b)$ (the least favorable null),

$$P\left(\hat{\theta}_1 - a \geqslant c_a - a \text{ or } \hat{\theta}_2 - b \leqslant c_b - b\right)$$

Electronic copy available at: https://ssrn.com/abstract=3508224

can be estimated by bootstrap:

$$P^*\left(\hat{\theta}_1^* - \hat{\theta}_1 \geqslant c_a - a \text{ or } \hat{\theta}_2^* - \hat{\theta}_2 \leqslant c_b - b\right)$$

The bootstrapped P-value is thus

$$P^*\left(\hat{\theta}_1^* - \hat{\theta}_1 \geqslant \hat{\theta}_1 - a \text{ or } \hat{\theta}_2^* - \hat{\theta}_2 \leqslant \hat{\theta}_2 - b\right) = P^*\left(\hat{\theta}_1^* \geqslant 2\hat{\theta}_1 - a \text{ or } \hat{\theta}_2^* \leqslant 2\hat{\theta}_2 - b\right)$$

The program to implement is then

$$\min_{b=g(a)} P^*\left(\hat{\theta}_1^* \geqslant 2\hat{\theta}_1 - a \text{ or } \hat{\theta}_2^* \leqslant 2\hat{\theta}_2 - b\right) \tag{9}$$

The machine is chosen if the minimized value is smaller than a given level. However, note that (9) is identical to

$$\max_{b=g(a)} P^*\left(\hat{\theta}_1^* \leqslant 2\hat{\theta}_1 - a \text{ and } \hat{\theta}_2^* \geqslant 2\hat{\theta}_2 - b\right)$$

which is exactly (8).

There is a large literature on testing multivariate equalities based on penalizing objective functions. Many of these testing procedures can be used to generate a program that optimizes the P-value along $a, b = g(a)$ to select the optimal point on the machine ROC.

2.4 Replacement of A Subset of Human Decision Makers

The Bayesian approach and frequentist approach translate into related algorithms for replacing a subset of the human decision makers. First, a Bayesian approach can compute (1) and replaces a human decision maker by the machine when (1) is larger than a given confidence level.

Second, both the Bayesian approach and the frequentist approach can provide confidence sets (also called credible sets in the Bayesian setting). These confidence set procedures also translate into algorithms for replacing a subset of the human decision makers. The following provides more details.

Conventional confidence and credibility levels suggest that we are conservative when making decisions to replace humans with machines. Specifically, only when there exists a point on the machine's ROC curve, which is better than any point in a confidence set of the human's FPR/TPR of level α, the machine's decision is considered to be better than the human's decision in an α confidence sense. In other words, only when a sufficiently large portion of the FPR/TPR joint distribution corresponding to α confidence level is below the machine's ROC curve, we will replace the human's decision by the machine's.

Electronic copy available at: https://ssrn.com/abstract=3508224

Next consider an illustration using a conventional oval-shaped confidence region based on the asymptotic normal distribution of the estimated human's FPR/TPR pairs. Note however that the following discussion applies to any valid confidence region of any shape.

Denote by $(HFPR_1, HTPR_H)$ the point that corresponds to the highest TPR on the oval, and $(HFPR_S, HTPR_2)$ the point that corresponds to the smallest FPR on the oval. The point $(HFPR_S, HTPR_H)$ corresponds to P in Figure 3. Regarding the position of the human's oval shape and machine's ROC curve, we have three cases.

1. Case 1: The human's oval area and the point $(HFPR_S, HTPR_H)$ are all below the ROC curve. In this case, the human decision maker is "worse" than the machine, and hence can be replaced by the machine. This case corresponds to Figure 3. In this case, one can find two points on the ROC curve $(RFPR_1, HTPR_H)$ and $(HFPR_S, RTPR_2)$ corresponding to points B and A in Figure 3. Any point from A to B corresponds to a better machine's decision.

2. Case 2: The entire human's oval confidence area is below the ROC curve, but the point $(HFPR_S, HTPR_H)$ is above the ROC curve. In this case, although different points on the machine ROC curve can achieve a better decision than different points in the confidence region of the human TPR/FPR pair, one cannot find a nonempty fraction on the machine's ROC curve that is better than *any* point in the human's confidence set so as to account for the randomness of the estimated human TPR/FPR pair. In this case, which corresponds to Figure 4, the human decision maker is not replaced by machine.

3. Case 3: The human's oval area has a certain area above the machine's ROC curve, therefore the human decision maker is not replaced by machine. It could be either because the human decision maker is sufficiently capable, or because his/her FPR/TPR pair is not measured precisely enough, for example due to a lack of historical data, which makes the confidence area (oval) quite large. This case corresponds to Figure 5.

In summary, if the P point $(HFPR_S, HTPR_H)$ is below the ROC curve, the human decision maker is replaced, otherwise he/she is not.

[Figure 4 about here.]

[Figure 5 about here.]

Given a decision to replace a human diagnosis by the machine learning algorithm, the remaining question of which point on the machine ROC curve to use can be addressed by, for example, (1) or (3).

Electronic copy available at: https://ssrn.com/abstract=3508224

3 Data and Algorithm Description

We illustrate our methodology using an empirical data set in which doctors need to make diagnosis on high risk pregnancies and birth defects. We first divide our entire sample into two sub-samples. The first subsample is used to find incapable doctors. Using the second subsample, we replace those incapable doctors by machine algorithms and use "machines" to make decisions for observations where the doctor has been replaced. This replacement is under the assumption that the less capable doctors completely follow the rules of machine algorithm in making decisions. We find that, the overall performance of the diagnose procedure is largely improved by combining capable doctors and machine algorithms.

The data of our study comes from the National Free Pre-Pregnancy Checkups (NFPC). Starting from 2010, this project offers free health checkup for couples planning a pregnancy and is conducted across 31 provinces in China. The data set contains more than 300 features for each observation, including age, demographic characteristics, results from medical examination and clinical test, disease and medication history, pregnancy history, as well as lifestyle and environmental information of both wife and husband. The data also has the true pregnancy outcome, which is denoted as normal ($y = 0$) or defect ($y = 1$). In addition, the data set records doctors' IDs and diagnosis of pregnancy risk. The doctor's diagnosis is graded in 4 levels: 0 for normal, 1 for high-risk of female, 2 for high-risk of male and 3 for high-risk for both female and male. In this paper, we re-group doctor's diagnosis into 2 levels, where a grade of 0 corresponds to normal pregnancy and any higher grade corresponds to a diagnosis of risky pregnancy.

The original dataset includes information about 3,330,304 couples that have pregnancy outcomes between January 1, 2014 and December 31, 2015. We exclude the samples with missing information on doctor's diagnosis and those for which more than 50% of feature values are missing. The final data set used for analysis includes 1,137,010 couples, who are diagnosed by 28,716 doctors. Of these observations, 61,184 couples (5.38%) have a defect birth outcome.

The basic statistical measurement of the quality of a binary classifier is accuracy, which is the proportion of correct predictions among the total number of cases examined. However, accuracy itself is not adequate for measuring prediction quality in our cases. As the adverse birth rate is about 5%, a naive classifier that categories all cases as low-risk would achieve an accuracy of nearly 95%. This is clearly controversial. The doctors' overall accuracy is at 73.63%, and 24.04% of couples are diagnosed as high-risk pregnancy. The false positive rate (the rate of misjudgment of normal birth as high-risk pregnancy) of all doctors is 0.2379. The true positive rate (TPR) of all doctors is 0.2843. In contrast, the FPR and TPR of the naive predictor are both zero. In other words, doctors are willing to tolerate a higher FPR

Electronic copy available at: https://ssrn.com/abstract=3508224

in order to achieve a higher TPR.

To improve the precision of statistical inference, we focus on doctors who diagnosed more than 300 patients, and we have a total of 584 such doctors corresponding to 584,181 cases. We first randomly split our sample to two parts. The first part is the "classification" part, i.e. we use the first part data to train a good machine algorithm, and then classify "capable" and "incapable" doctors based on the relative performance between machines and doctors. The second part is the "performance" part, i.e. where we utilize machines to replace incapable doctors in decision making. In particular, we split each doctor's cases into two parts for "classification" and "performance", with a respective ratio of 7:3, using stratified splitting methods. The random splitting process is done through each class of data (all cases are grouped into four classes as true positive cases, true negative cases, false negative cases and false positive cases) with the same ratio 7:3. Then data in each class is merged together for both "classification" and "performance" parts. There are, in total, 408,661 cases in the classification sample and 175,520 cases in the performance sample.

The first objective in the classification exercise is to train a machine algorithm and evaluate its performance. We further split our classification sample into two parts: The first part is the training sample used to estimate the parameters in a machine algorithm; The second part is the validation sample that can be used to evaluate the performance of machine algorithm and obtain machine's ROC curve. We split each doctor's cases in the classification sample into the two parts for training and validation with a ratio of 4:3, using the same stratified splitting methods as described above. There are 233,435 cases in training sample and 175,226 cases in validation sample.

We experimented with several machine learning algorithm, and noted that that decision trees are widely used in many machine learning applications. However, a single tree method usually has high variance in spite of its low bias. It can be easily affected by even small amount of noises in the data, and training it tends to overfit the data. In contrast, random forest (RF) is a well-known and commonly used ensemble learning algorithm proposed by Breiman (2001). It overcomes these problems by constructing a collection of decision trees that are trained using different feature subspaces with bootstrapped samples. The predictions of each tree are aggregated as output to make predictions. A more detailed description of the random forest algorithm is provided in Appendix A.4.[6]

[6]For the parameters setting, we choose the number of estimators $N = 100$, the number of max features per node $M = 50$ and the minimum number of samples required to split is set to 50.

Electronic copy available at: https://ssrn.com/abstract=3508224

4 Empirical Results for AI Assisted Doctors

Next we present the empirical results from three different approaches: (1) the heuristic approach described in section 2.4 combined with random sampling of points on the dominating segment of the machine ROC; (2) the Bayesian approach of section 2.2; (3) the frequentist approaches in (3). We have not yet implemented all the theoretical discussions but merely focus on a subset for illustration.

4.1 Asymptotic Confidence Set with Random ROC Sampling

We first implement section 2.4 combined with random sampling of points on the dominating segment of the machine ROC. For each doctor, we obtain 100 bootstrap samples from his/hers historical records, and obtain the 2 sigma (95.4%) confidence oval area of (FPR, TPR) pairs using the bootstrapped covariance matrix of FPR and TPR centered around the sample estimates. Then we find the highest TPR on the oval i.e. the point $(HFPR_1, HTPR_H)$, and the smallest FPR on the oval, i.e. the point $(HFPR_S, HTPR_2)$. Next we obtain the point $P(HFPR_S, HTPR_H)$ as figure 3 shows, as the representative point of this doctor. Then we can classify doctors into two groups.

1. Incapable doctors. The doctor's representative point $(HFPR_S, HTPR_H)$ is below machine's ROC curve, which corresponds to case 1 in section 2.4. These doctors are replaced in future decision making.

2. Other doctors. The doctor's representative point $(HFPR_S, HTPR_H)$ is above machine's ROC curve, which corresponds to case 2 and case 3 in section 2.4. These doctors are either more capable than machines or their capability can not be precisely measured likely due to small amount of data (resulting in a larger size confidence set area). They hence are not replaced by the machine algorithm in future decision making.

After we search for incapable doctors applying this algorithm to the "classification" sample, we find that among the 584 doctors, 372 incapable doctors are replaced by machines Therefore, we have 372 "machine" doctors (64%) and 212 (36%) human doctors ready for diagnosing patient cases in the "performance" sample.

4.1.1 Performance Evaluation

We evaluate the overall performance of combining the decisions of capable doctors and machine algorithm on performance sample by calculating the overall get the FPR and

Electronic copy available at: https://ssrn.com/abstract=3508224

TPR. For doctors who are not in the incapable group, their decisions are not replaced by machines, so we use their origin decisions in the evaluation process.

For each of incapable doctors, we need to determine a threshold value c for the machine algorithm to make classification decisions. Using the point $P(HFPR_S, HTPR_H)$ for each of these doctors, we find two points on machine's ROC curve, i.e. the point $(HFPR_S, RTPR_1)$ and the point $(RFPR_2, HTPR_H)$, which correspond to points A and B in figure 3. As in Lemma 2.1, the fraction of points B to A on machine's ROC curve offers a set of better decision rules than $(HFPR_S, HTPR_H)$ does, in the sense of achieving higher TPRs and lower FPRs. Denote by c_2 the decision threshold that corresponds to point $A(HFPR_S, RTPR_1)$ and by c_1 the decision threshold corresponding to point $B(RFPR_2, HTPR_H)$ on the machine's ROC curve. As both TPR and FPR are monotonically decreasing functions of the threshold, we thus have $c_1 < c_2$. The thresholds interval $[c_1, c_2]$ can be interpreted as a "better" range where the machine would outperform the doctor in terms of TPR/FPR comparison if the machine's decision threshold c is chosen along this interval. The threshold values depend on the identity of each doctor who is to be replaced by the machine algorithm, and are more accurately denoted as $[c_{d_i,1}, c_{d_i,2}]$.

Our first strategy is to visualize the set of TPR/FPR pairs when the threshold for classification used by the machine algorithm in the machine and capable doctor combination ranges over the entire interval $[c_{d_i,1}, c_{d_i,2}]$. for each doctor. We define this set as the overall performance curve, i.e. a set of FPR/TPR pairs where machine's threshold ranges from $c_{d_i,1}$ to $c_{d_i,2}$ for each incapable doctor. This is numerically implemented using the following uniformly sampling algorithms (N is the number of points sample on the curve).

1. For each incapable doctor d_i, set the step of threshold $c_{d_i,step} = (c_{d_i,2} - c_{d_i,1})/N$
2. For $j = 0$ to N:
   (a) For each incapable doctor d_i:
   The threshold of d_i is set to $c_{d_i} = c_{d_i,1} + c_{d_i,step} * j$. Random forest model makes prediction $\widehat{p}(x)$ on each case x of d_i in the testing data set, and makes decision using the specific threshold c_{d_i}: $\widehat{D}(x) = \mathbb{1}(\widehat{p}(x) > c_{d_i})$.
   (b) For other doctors, use their origin decisions on the testing data set.

   Obtain a single FPR/TPR pair result T_j by combining the decisions of (a) and (b).
3. Collect the uniformly sampled FPR/TPR pairs $\{T_j\}_{j=1}^{N}$.

The overall experiment procedure is summarized in Appendix A.5. In this experiment, we sample $N = 20$ points on thresholds interval.

Electronic copy available at: https://ssrn.com/abstract=3508224

Figure 6 shows the results of this experiment. The aggregated FPR/TPR pair of all doctors on testing set is at 0.2065 for FPR and 0.2264 for TPR, which corresponds to the blue point in figure 6. The ROC curve of random forest model on the testing set is the green curve in that figure, and the area under curve (AUC) of the model is 0.6834. The yellow point and cyan interval are results of overall performance by replacing "incapable" doctors with machine decisions on the testing set. Any point on the cyan interval (including the yellow point) has lower FPR but higher TPR compared to the blue "all doctor" FPR/TPR pair, representing better performance than the doctors overall. In particular, the yellow point represents the case when the threshold of each incapable doctor d_i is set to $c_{d_i,1}$ as section 4.1.1 describes, which is one endpoint of interval that has the highest FPR and highest TPR. The yellow point has 0.1770 of FPR and 0.3201 of TPR. It means that by replacing "incapable" doctors with machines, we improve the TPR by 41.4%, while at the same time also reduce by 14.3% the FPR comparing to the case where the diagnoses are all done by the doctors.

Another endpoint of the cyan interval, i.e. the lower left point on the interval, achieves 0.1352 of FPR and 0.2731 of TPR. If we use the thresholds corresponding to this point as machine's thresholds, we would increase by 20.6% in TPR and reduce by 34.5% in FPR comparing to all doctors diagnosis. Compared to the yellow point, there is less improvement in the TPR. But this is compensated by a larger reduction in the FPR, meaning that there is a smaller amount of type I error. Essentially, the cyan interval merely reflects a trade-off between improving TPR and reducing FPR, but any point along the interval works better than humans do. Furthermore, this endpoint (the red point on figure) is also the point that achieves the maximum F1 score among the cyan interval curve, which achieves a F1 score of 0.144.

[Figure 6 about here.]

4.1.2 Robustness Checks

To check the robustness of our methods on AI assisted decision making, we conduct the experiments again based on doctors who diagnosed more then 500 patients. Other settings in the experiment are kept the same. We have 367 such doctors and 495,320 cases totally. For the "classification" data set, we have 197,978 cases in the training sample and 148,579 cases in the validation sample. Altogether 346,557 cases are used as historical records for finding incapable doctors. For the "performance" data set, we have 148,763 cases for evaluating the performance of the combination of human doctors and "machine" doctors. After the classification process in section 4, we find 276 incapable doctors among 367 doctors. In

Electronic copy available at: https://ssrn.com/abstract=3508224

other words, 75% of the doctors would be replaced by machines and 25% of human doctors would be retained in the "performance" part.

[Figure 7 about here.]

Figure 7 shows the results of this experiment. Similar to the remarks after figure 6, the aggregated FPR/TPR pair of all doctors on the testing set (blue point) is 0.1942 for FPR and 0.2135 for TPR. The AUC of the RF model is 0.6920. The yellow point on the cyan interval is 0.1682 in FPR and 0.3417 in TPR. This represents an improvement of 60.0% in TPR for the experiment, and a corresponding reduction by 13.4% of FPR. The lower left point of the cyan interval achieves 0.1203 of FPR and 0.2766 of TPR. If we use the thresholds corresponding to this point as the machine's decision rule, we would obtain an improvement of 29.6% in TPR and a reduction by 38.1% of FPR comparing to all doctors diagnoses. The point that achieves the maximum F1 score among the cyan curve is the red point in figure 7, which is close to the lower left point and registers a F1 score of 0.154. The overall performance of machine with capable doctors does not differ substantially from is not much the experiments conducted in section 4.1.1, providing evidence for the robustness of our methods.

The above experiments demonstrate the potential capability and robustness of our algorithms at assisting doctors with the diagnosis of risky pregnancy. Using machines to replace incapable doctors, the overall performance of the diagnosis is drastically improved. We also provides an interval for users to choose from to balance the trade-off between improving TPR and reducing FPR. At any point on the interval, the overall performance of the combined diagnosis is better than that by the humans.

4.2 Results of Bayesian Approach

For each doctor D_i and the doctor's historical data $\{(X_i, \widehat{Y}_i, Y_i)\}$, where $\widehat{Y}_i$ is doctor's diagnosis and Y_i is the ground-truth label for case X_i, we can use three parameters to specify the doctor's model, i.e. $p_1 = \mathbb{E}Y\widehat{Y}, p_2 = \mathbb{E}(1-Y)\widehat{Y}, p_3 = \mathbb{E}Y(1-\widehat{Y})$ (we also have $p_4 = \mathbb{E}(1-Y)(1-\widehat{Y})$. However $p_1 + p_2 + p_3 + p_4 = 1$ so p_4 is not a free parameter).

Notice we have

$$\text{FPR} = \frac{\mathbb{E}(1-Y)\widehat{Y}}{\mathbb{E}(1-Y)}, \qquad \text{TPR} = \frac{\mathbb{E}Y\widehat{Y}}{\mathbb{E}Y},$$

we also denote $\theta_1 = \text{FPR}$, $\theta_2 = \text{TPR}$, and we can get the formula $\theta = h(p)$ shows the relation between θ and p, i.e.

$$\theta_1 = \frac{p_2}{p_2 + p_4}, \theta_2 = \frac{p_1}{p_1 + p_3}.$$

Electronic copy available at: https://ssrn.com/abstract=3508224

We first set a prior $\pi(p)$, which is a symmetry Dirichlet distribution with four parameters $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, and we have $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha = 0.01$. So $\pi(p)$ is a uniform prior on p_1, p_2, p_3, p_4.

We group the doctor's historical cases into four class according to the values of $\widehat{Y}$ and Y of each case, and get the frequency of each class, i.e.

1. $n\widehat{p}_1$ equals to the number of cases that $Y = 1, \widehat{Y} = 1$;
2. $n\widehat{p}_2$ equals to the number of cases that $Y = 0, \widehat{Y} = 1$;
3. $n\widehat{p}_3$ equals to the number of cases that $Y = 1, \widehat{Y} = 0$;
4. $n\widehat{p}_4$ equals to the number of cases that $Y = 0, \widehat{Y} = 0$.

Since the data follows a multinomial distribution with four categories, given p, the data likelihood is

$$L(Y|p) = \binom{n}{n\widehat{p}_1} p_1^{n\widehat{p}_1} \binom{n - n\widehat{p}_1}{n\widehat{p}_2} p_2^{n\widehat{p}_2} \binom{n - n\widehat{p}_1 - n\widehat{p}_2}{n\widehat{p}_3} p_3^{n\widehat{p}_3} p_4^{n\widehat{p}_4}.$$

We combine the prior distribution and the data likelihood to analytically compute the posterior distribution, which is also a Dirichlet distribution with parameters $(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4)$, where

$$\widehat{\alpha}_i = \alpha + n\widehat{p}_i = 0.01 + n\widehat{p}_i,$$

for $i = 1, 2, 3, 4$.

The posterior distribution of $\theta = (\theta_1, \theta_2)$ can be simulated. First, draw S samples of $\{p_i = (p_{1i}, p_{2i}, p_{3i}, p_{4i})\}_{i=1}^S$ from the posterior Dirichlet distribution with parameters $(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4)$. After that, we can get the $\{\theta_i = (\theta_{1i}, \theta_{2i})\}_{i=1}^S$ with the simulated S samples of p using the formula $\theta = h(p)$ described above.

The machine's ROC curve can be represented as the formula $\theta_2 = g(\theta_1)$. For any point $P(a, g(a))$ on machine's ROC Curve, we can obtain the portion that P dominate $\theta_r \in \{\theta_i\}_{i=1}^S$, which is defined as

$$q\,(a) = \frac{1}{S} \sum_{i=1}^{S} \mathbb{1}(\theta_{1i} \geqslant a, \theta_{2i} \leqslant g(a)),$$

and we can find the maximum of q on the ROC curve, i.e.

$$q_{max} = \max_{a \in (0,1)} \left[\frac{1}{S} \sum_{i=1}^{S} \mathbb{1}(\theta_{1i} \geqslant a, \theta_{2i} \leqslant g(a)) \right],$$

Electronic copy available at: https://ssrn.com/abstract=3508224

using numerical methods. If $q_{max} \geqslant 0.95$, we mark the doctor D as an incapable doctor, who would be replaced by machines in future decisions, and the threshold c that corresponds to the point which maximizes q on ROC is used as decision threshold for machine algorithms. Otherwise, the doctor would not be replaced by machines.

For example, figure 8 shows the case where a doctor would be replaced by machine. The maximum coverage rate q achieves 0.970 at the black point on ROC. However, in figure 9, the maximum coverage rate only achieves 0.279, so there is not sufficient evidence to replace the doctor's decisions with machine's.

[Figure 8 about here.]

[Figure 9 about here.]

Figure 10 (a) shows the result of experiments on doctors who diagnosed more than 300 cases. A total of 255 out of 584 doctors are classified as incapable doctors. So we have 329 (56%) human doctors and 255 (44%) "machine" doctors. Overall, human doctors achieve 0.2065 of FPR and 0.2264 of TPR, which corresponds to the blue point on the figure. By replacing incapable doctors' decisions with machine decisions, the combined decision results in 0.1871 of FPR and 0.3255 of TPR, i.e. the yellow point on the figure, which indicates an increase of 43.8% in TPR and a reduction of 9.39% in FPR.

Figure 10 (b) shows the result of experimenting with on doctors who diagnosed more than 500 cases. Among all 367 doctors, 203 of them are classified as incapable doctors, who account for 55% of the total. The overall performance of human doctors is at 0.1942 of FPR and 0.2135 of TPR. By replacing the decisions made by incapable doctors with machine decisions, the overall FPR is at 0.1792 and TPR is at 0.3332. So we obtain an improvement of 56.1% on TPR and a reduction of 7.68% on FPR.

[Figure 10 about here.]

4.3 Results of Frequentist Approach

4.3.1 First Approach

This subsection implements (3). For each doctor D, we obtain S bootstrapped samples of the doctor's diagnosis cases, and get the θ values for the S samples i.e. $\{\theta_i = (\theta_{1i}, \theta_{2i})\}_{i=1}^{S}$. Given machine's ROC curve $\theta_2 = g(\theta_1)$, for an arbitrary point $(a, g(a))$ on ROC, we can use (2) to estimate the probability that the point $(a, g(a))$ dominates doctor's θ, i.e,

$$P(\theta_1 \geqslant a, \theta_2 \leqslant g(a)).$$

Electronic copy available at: https://ssrn.com/abstract=3508224

Based on this, we want to obtain the maximum of probability that

$$\max_{a\in(0,1)} P(\theta_1 \geqslant a, \theta_2 \leqslant g(a))$$

on machine's ROC curve. If the maximized value is larger than the given confidence level (95%), the doctor D is classified as incapable doctor, and the threshold value corresponded to the maximized pair of $(a, g(a))$ on machine's ROC is chosen as decision threshold. Otherwise we would not replace doctor's decisions with machine results.

Figure 11 (a) shows the result of this experiment under frequentist approach. Among all 584 doctors who diagnosed more than 300 cases, 261 of them are marked as incapable doctors. So we have 323 (55%) human doctors and 261 (45%) "machine" doctors totally. Overall, human doctors achieve 0.2065 of FPR and 0.2264 of TPR, as the blue point shown on the figure. By replacing incapable doctors' decisions with machine decisions, the combined decision making results in 0.1885 of FPR and 0.3226 of TPR, i.e. the yellow point on the figure, which indicates an increase of 42.5% in TPR and a reduction of 8.73% in FPR.

Figure 11 (b) shows the result of experiments on doctors who diagnosed more than 500 cases. Among all 367 doctors, 203 of them are classified as incapable doctors, who account for 55% of the total. The overall performance of human doctors is at 0.1942 of FPR and 0.2135 of TPR. By replacing the decisions made from incapable doctors with machine decisions, the overall FPR is at 0.1732 and TPR is at 0.3252. So we obtain an improvement of 52.3% on TPR and a reduction of 10.8% on FPR.

[Figure 11 about here.]

4.3.2 Two Alternative Approaches

This subsection implements two alternative frequentist approaches, i.e. (4) and (5). Firstly, we would present the results of approach (4). For each doctor D, given machine's ROC $\theta_2 = g(\theta_1)$ and D's FPR/TPR pair $\left(\hat{\theta}_1, \hat{\theta}_2\right)$, we would find a pair (a, b) that maximize the estimated probability of machine algorithm outperforming human doctor D, i.e.

$$\max_{a\in\left(0,\hat{\theta}_1-g^{-1}\left(\hat{\theta}_2\right)\right),b=g\left(\hat{\theta}_1-a\right)-\hat{\theta}_2} P^*\left(\hat{\theta}_1^* - \hat{\theta}_1 \leqslant a,\ \hat{\theta}_2^* - \hat{\theta}_2 \geqslant -b\right),$$

where the constrains guarantee the pair $\left(\hat{\theta}_1 - a, \hat{\theta}_2 + b\right)$ lies on machine ROC, and $\left(\hat{\theta}_1^*, \hat{\theta}_2^*\right)$ is obtained from the bootstrapped samples of doctor $D's$ diagnosis cases. If the maximized value is greater than the given confidence level, the doctor D would be classified as incapable doctor, and the threshold value of point $\left(\hat{\theta}_1 - a, \hat{\theta}_2 + b\right)$ on machine ROC with (a, b)

Electronic copy available at: https://ssrn.com/abstract=3508224

corresponding to the maximized P^* is chosen as decision threshold. Otherwise we would not replace doctor's decisions with machine results.

Figure 12 shows the results of experiments under this frequentist approach. We choose 95% as confidence level, and bootstrapped 1000 times for each doctor. Figure 12 (a) shows the result of experiments on doctors who diagnosed more then 300 cases. Of all 584 doctors, 273 of them are labeled as incapable doctors. So, we would have 311 (53%) human doctors and 273 (47%) "machine" doctors take part in the evaluation of testing data set. The human doctors achieve 0.2065 of FPR and 0.2264 of TPR in the whole testing data set, as the blue pair shows on the graph. By replacing incapable doctors' decisions with machine's decision, the overall FPR is at 0.1939 (a reduction of 6.12%) and the TPR is at 0.3410 (an increase of 50.6%), showed by the yellow pair.

Figure 12 (b) shows the result of experiments on doctors who diagnosed more then 500 cases. After classification, 211 of 367 doctors are marked as incapable doctors, so the "machine" doctors account for 57%. The blue pair in the graph is the FPR/TPR pair of overall doctors, which is at 0.1942/0.2135. After the replacement of incapable doctors' decisions, we would get the FPR at 0.1822 and the TPR at 0.3478, as the yellow pair shows. So, in this case, we get an improvement of 62.9% at TPR and a reduction of 6.18% at FPR.

[Figure 12 about here.]

The last is implementation of (5). Similar to approach (4), for each doctor D, given machine's ROC $\theta_2 = g(\theta_1)$ and D's FPR/TPR pair $\left(\hat{\theta}_1, \hat{\theta}_2\right)$, we would find a pair (a, b) that minimize the estimated probability of human doctor D superior to machine algorithm, i.e.

$$\max_{a\in\left(0,\hat{\theta}_1-g^{-1}\left(\hat{\theta}_2\right)\right),b=g\left(\hat{\theta}_1-a\right)-\hat{\theta}_2} P^*\left(\hat{\theta}_1^* - \hat{\theta}_1 \geqslant a,\ \hat{\theta}_2^* - \hat{\theta}_2 \leqslant -b\right),$$

where the constrains guarantee $\left(\hat{\theta}_1 - a, \hat{\theta}_2 + b\right)$ lies on machine ROC, and $\left(\hat{\theta}_1^*, \hat{\theta}_2^*\right)$ is calculated from the bootstrapped samples of doctor $D's$ diagnosis cases. If the minimized value is less than one minus the given confidence level, the doctor D would be classified as incapable doctor, and the threshold value of point $\left(\hat{\theta}_1 - a, \hat{\theta}_2 + b\right)$ on machine ROC with (a, b) corresponding to the minimized P^* is chosen as decision threshold. Otherwise doctor D's decisions would not be replaced.

Figure 13 shows the results of experiments under this frequentist approach. We choose 95% as confidence level as before, so one minus the confidence level would be 5%. And the number of bootstrap time is also 1000 for each doctor. Figure 13 (a) shows the result of experiments on doctors who diagnosed more then 300 cases. Among all 584 doctors, 522 of them are labeled as incapable doctors. So, we would have only 62 (11%) human doctors and

Electronic copy available at: https://ssrn.com/abstract=3508224

522 (89%) "machine" doctors. Overall, human doctors achieve 0.2065 of FPR and 0.2264 of TPR in testing data set, as the blue pair shows on the graph. By replacing incapable doctors' decisions with machine's decision, the overall FPR is at 0.1963 (a reduction of 4.96%) and the TPR is at 0.3813 (an increase of 68.4%), showed by the yellow pair.

Figure 13 (b) shows the result of experiments on doctors who diagnosed more then 500 cases. After classification, 350 of 367 doctors are marked as incapable doctors, so the "machine" doctors account for 95%. The blue pair in the graph is the FPR/TPR pair of overall doctors, which is at 0.1942/0.2135. After the replacement of incapable doctors' decisions, we would get the FPR at 0.1851 and the TPR at 0.3813, as the yellow pair shows. So, in this case, we get an improvement of 78.6% at TPR and a reduction of 4.67% at FPR.

[Figure 13 about here.]

It seems that approach (5) would label more incapable doctors than (4), and this may be due to non-equivalent selection of confidence level. In implementation of (4), the confidence level is set to 95%. In implementation of (5), we used $(1 - 95\%)$ as threshold for judging incapable doctors. This implies a condition that the probability of machine definitely outperforming human decision maker and the probability of human definitely superior to machine are add up to 1. However, this may not be true in the estimation of doctor and machine in our experiments, and there're cases that we cannot judge definitely whether human or machine performs better, for example, when machine's FPR and TPR are both lower then human's FPR and TPR.

4.4 Geographic Characteristics of Replaced Doctors

The approaches mentioned above provide methods for filtering incapable doctors that have poorer diagnostic capability than machines. We create a dummy variable "incapable" that is set to 1 when a doctor is replaced by machine, and 0 otherwise. The next step is to check whether locations of doctors have significant influence on this replacement. Particularly, we choose two factors: the first factor is the provincial Gross Domestic Product (GDP) from China City Statistical Yearbook of year 2014, which coincides with the starting year of our data. The other factor is a dummy variable that is set to 1 when the doctor is in a township (lower level) clinic, and 0 when he/she is in an urban (higher level) clinic/hospital.

We conduct a regression analysis of the dummy variable "incapable" on these two factors for all doctors who diagnosed more than 300 cases, excluding the ones that don't have records on the township factor. Table 1 shows the results of linear and logistic regression using both Frequentist and Bayesian approaches. Doctors from township clinic have significantly (10%) higher probability to be replaced by machines relative to those from urban clinics. Particularly, with the Bayesian approach, 50.5% of doctors from township clinic are

Electronic copy available at: https://ssrn.com/abstract=3508224

replaced by machine, while 40.2% of urban clinic doctors are replaced. Using Frequentist method, we find 50.2% and 39.3% of township and urban clinic doctors are replaced by machines, respectively. The provincial GDP has a negative coefficient in the regression, which informs doctors from low GDP province tend to be replaced; however, the statistical insignificance informs a less strength of GDP in detecting the capability of doctors compared to the township dummy.

Overall, these results inform that artificial-intelligence technology tends to benefit rural area more than developed cities in decision making. To the best of our knowledge, our result is the *first* to empirically show that the decision making of artificial intelligence is more beneficial to poor areas relatively to developed areas.[7] This is easy to understand: poor areas normally have more less-educated and lower-skilled people who are lack of sufficient knowledge in making a good decision. This has an important implication for poverty reduction with high-tech.

[Table 1 about here.]

5 Conclusion

In this paper, we propose our principle of how artificial intelligence can assisted human in decision making. First, by assuming that preference of individual decision maker is constant, we can make comparison on the ability of decision making between human and machine, and identify the less capable decision makers. Second, given that the incapable human decision maker will completely follow the decision of machines (or replaced by machines), we propose decision rules based on the machine's ROC, including randomization and maximizing the level of statistical confidence.

By replacing these less capable decision makers with machines, we would obtain improved performance in the quality of future decisions. Experiments from a data set of the Pre-Pregnancy Checkups shows that by replacing the less capable doctors, the overall performance of risky pregnancy detection achieves much improvement. Furthermore, less capable doctors are more likely to come from rural clinics than urban ones. This indicates that artificial intelligence is more beneficial in under-developed areas.

Our current paper addresses the issue of sampling uncertainty in a simple setting with an individual doctor. There are many open questions and directions of extensions. For example, empirically we are likely to be testing multiple doctors. There is a large literature

[7] In the non-decision-making area, there're are related literature. For example, Reis et al. (2004) developed a fuzzy-logic based system to predict the need of neonatal resuscitation for areas with poor medical resources, Onu et al. (2017) proposed a cheap and easy-to-use machine-learning based system for detecting asphyxia.

Electronic copy available at: https://ssrn.com/abstract=3508224

on multiple hypothesis testing that includes the use of Bonferroni corrections, stepwise procedures, and other procedures that take the joint distribution across doctors into account, in order to control family-wise error rates. These are issues for future research.

A Appendix

A.1 Incentive-feature Dependence

In this paper, we assume away the incentive-feature dependence (as shown in Assumption 1), so that we are able to make comparison between machines and individual decision makers. However, each individual decision maker may not have a constant preference, and may change her cutoff threshold c based on observed features. In this case, each decision maker does not need to be represented by a single point on the ROC curve. For example, Chandra and Staiger (2011) find that hospitals treat similar patients differently due to consideration of commercial benefits. Therefore c depends on both hospitals and the location of patients.

If this type of incentive heterogeneity exists across the sample of decision makers, or within each individual, the aggregate TPR/FPR pair can lie below the optimal ROC curve, or even below the 45 degree line, even if humans have better information processing capacities than machines. Lemma A.1 provides a formal proof.

Lemma A.1. *In the absence of information heterogeneity, if the cutoff threshold c varies among decision makers or within a single decision maker (incentive heterogeneity), the aggregate PTPR/PFPR pair of decision makers is below the optimal ROC curve.*

Proof. To prove this, we write the pair as

$$\begin{aligned}\text{PFPR} =& \frac{1}{1-p}\int \lambda(x)\left(1-p(x)\right) f(x)\,\mathrm{d}x,\\ \text{PTPR} =& \frac{1}{p}\int \lambda(x)\, p(x)\, f(x)\,\mathrm{d}x,\end{aligned}$$

where $\lambda(x) = \int \mathbb{1}\left(p(x) > h(x,v)\right) f(v)\,\mathrm{d}v$, note that the cutoff variable $h(x,v)$ depends on x and a random variable v. We assume that the decision rule $\hat{Y} = \mathbb{1}\left(p(x) > h(x,v)\right)$ has some classification ability, i.e. $\mathbb{1}\left(p(x) > h(x,v)\right) \not\equiv 0$ and $\mathbb{1}\left(p(x) > h(x,v)\right) \not\equiv 1$, therefore, $0 < \lambda(x) < 1$. Then for a c^* satisfying

$$\alpha(c^*) = \int \mathbb{1}\left(p(x) > c^*\right)\left(1-p(x)\right) f(x)\,\mathrm{d}x = \text{PFPR} = \int \lambda(x)\left(1-p(x)\right) f(x)\,\mathrm{d}x \quad (10)$$

Electronic copy available at: https://ssrn.com/abstract=3508224

on the optimal ROC curve, given the FPR $\alpha(c^*)$, we can find the TPR $\beta(c^*)$:

$$\beta(c^*) = \int \mathbb{1}(p(x) > c^*)\, p(x)\, f(x)\, \mathrm{d}x. \tag{11}$$

Since $\alpha(c^*)$ and $\beta(c^*)$ are on the optimal ROC curve, by the Neyman-Pearson Lemma, there must be some positive η_1 and η_2 such that $\mathbb{1}(p(x) > c^*)$ (but not $\lambda(x)$) solves

$$\arg\max_{\phi(\cdot)} \eta_1 \int \phi(x)\, p(x)\, f(x)\, \mathrm{d}x - \eta_2 \int \phi(x)\,(1 - p(x))\, f(x)\, \mathrm{d}x. \tag{12}$$

Hence $\eta_1\beta(c^*) - \eta_2\alpha(c^*) > \eta_1\mathrm{PTPR} - \eta_2\mathrm{PFPR}$, and thus $\beta(c^*) > \mathrm{PTPR}$. ■

A.2 Proof of Lemma 2.1

As shown in the population model of Feng et al. (2019), the Neyman Pearson lemma shows that a decision rule of $D = \mathbb{1}(p(x) > c)$ is the same as the maximization of a utility function of linear combination of TPR and FPR:

$$\phi\mathrm{TPR} - \eta\mathrm{FPR}$$

where $\phi = pC_{1A} > 0$ and $\eta = (1-p)C_{0R} > 0$, and p is the probability of the positive outcomes. C_{1A} is the cost of wrong decision when the real outcome is positive and C_{0R} is the cost of wrong decision when the real outcome is negative. C_{1A} and C_{0R} are assumed to be constant for individual decision makers, but may not be known by researchers.

If a human's FPR/TPR pair $(\mathrm{HFPR}_0, \mathrm{HTPR}_0)$ is below the ROC curve, we then can find two points on the ROC curve $(\mathrm{HFPR}_0, \mathrm{RTPR}_0)$ and $(\mathrm{RFPR}_0, \mathrm{HTPR}_0)$, which correspond to A and B in Figure 2. Any point $(RFPR, RTPR)$ between $(\mathrm{RFPR}_0, \mathrm{HTPR}_0)$ and $(\mathrm{HFPR}_0, \mathrm{RTPR}_0)$ on the machine's ROC curve has a larger TPR than HTPR_0 and a smaller FPR than HFPR_0, i.e. $RFPR < \mathrm{HFPR}_0$ and $RTPR > \mathrm{HTPR}_0$. We hence have:

$$\phi RTPR - \eta RFPR > \phi\mathrm{HTPR}_0 - \eta\mathrm{HFPR}_0$$

Therefore, any point between $(\mathrm{RFPR}_0, \mathrm{HTPR}_0)$ and $(\mathrm{HFPR}_0, \mathrm{RTPR}_0)$ on the ROC curve has a larger utility and hence corresponds to a better decision than the human.

When the true propensity score function $p(x)$ has a continuous distribution supported on $(0, 1)$ in the population model of Feng et al. (2019), the limit ROC curve is continuous and concave, and closely approximates the finite sample ROC when the sample size is sufficiently large. However, with any given sample size, x_i is discretely distributed, and the ROC corresponds to a maximum of n discrete points instead of a connected curve, as shown

Electronic copy available at: https://ssrn.com/abstract=3508224

in figure

[Figure 14 about here.]

which plots the pair of, with $p_i = p(x_i)$

$$\text{TPR} = \frac{1}{\hat{p}} \sum_{i=1}^{n} y_i \mathbb{1}(p_i \geqslant c), \quad \text{FPR} = \frac{1}{1-\hat{p}} \sum_{i=1}^{n} (1 - Y_i) \mathbb{1}(p_i \geqslant c).$$

There are different ways to connect these points into a *curve*. Examples are

[Figure 15 about here.]

and

[Figure 16 about here.]

Suppose the machine algorithm discovers the true population propensity score $p(x)$, and human decisions are based on the same set of features x, in any finite sample (for example in the testing or validation sample) there is no guarantee that the human aggregate TPR/FPR pair will lie below the finite sample machine ROC that are formed based on a finite number of points.

In one special case we can provide a definitive finite sample comparison. Suppose the features X are discretely distributed with a finite number of support points: $x \in (x_j, j = 1, \ldots, J)$. The sample is used both for machine estimation and ROC computation. Then let $n_j = \sum_{i=1}^{n} \mathbb{1}(X_i = x_j)$, there is

$$p_j = \hat{p}(x_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} y_i$$

The machine ROC is then based on the pairs:

$$\begin{aligned} \text{mTPR}(c) = & \quad \tfrac{1}{\hat{p}} \textstyle\sum_{i=1}^{n} y_i \mathbb{1}(p_i \geqslant c) = \tfrac{1}{\hat{p}} \sum_{j=1}^{J} n_j p_j \mathbb{1}(p_j \geqslant c) \\ \text{mFPR}(c) = & \quad \tfrac{1}{1-\hat{p}} \textstyle\sum_{i=1}^{n} (1 - y_i) \mathbb{1}(p_i \geqslant c) = \tfrac{1}{1-\hat{p}} \sum_{j=1}^{J} (1 - p_j) n_j \mathbb{1}(p_j \geqslant c). \end{aligned}$$

These pairs only change when c crosses one of the p_j's. The following figure depicts such (points of) ROC

[Figure 17 about here.]

Suppose a human decision rule is $\hat{y}_i = \mathbb{1}(q(x_i) \geqslant c_h) \equiv \mathbb{1}(q_i \geqslant c_h)$ for some c_h, generating a pair of

$$\begin{aligned} \text{hTPR} = & \quad \tfrac{1}{\hat{p}} \textstyle\sum_{i=1}^{n} y_i \mathbb{1}(q_i \geqslant c_h) = \tfrac{1}{\hat{p}} \sum_{j=1}^{J} n_j p_j \mathbb{1}(q_j \geqslant c_h) \\ \text{hFPR} = & \quad \tfrac{1}{1-\hat{p}} \textstyle\sum_{i=1}^{n} (1 - y_i) \mathbb{1}(q_i \geqslant c_h) = \tfrac{1}{1-\hat{p}} \sum_{j=1}^{J} (1 - p_j) n_j \mathbb{1}(q_j \geqslant c_h). \end{aligned}$$

Electronic copy available at: https://ssrn.com/abstract=3508224

Then by a discrete version of the Neyman-Pearson argument, the pair $(hFPR, hTPR)$ must lie in the shaped area of Figure 17. To see this, suppose the contrary. Then for some c_m, there are both

$$\begin{aligned}\textstyle\sum_{j=1}^{J} n_j p_j \mathbb{1}\left(q_j \geqslant c_h\right) &\geqslant \textstyle\sum_{j=1}^{J} n_j p_j \mathbb{1}\left(p_j \geqslant c_m\right) \\ \textstyle\sum_{j=1}^{J} \left(1-p_j\right) n_j \mathbb{1}\left(q_j \geqslant c_h\right) &\leqslant \textstyle\sum_{j=1}^{J} \left(1-p_j\right) n_j \mathbb{1}\left(p_j \geqslant c_m\right).\end{aligned}$$

Or (with at least one inequality being strict)

$$\begin{aligned}\textstyle\sum_{j=1}^{J} n_j p_j \mathbb{1}\left(q_j \geqslant c_h\right) &\geqslant \textstyle\sum_{j=1}^{J} n_j p_j \mathbb{1}\left(p_j \geqslant c_m\right) \\ \textstyle\sum_{j=1}^{J} \left(1-p_j\right) n_j \mathbb{1}\left(q_j < c_h\right) &\geqslant \textstyle\sum_{j=1}^{J} \left(1-p_j\right) n_j \mathbb{1}\left(p_j < c_m\right).\end{aligned}$$

Taking a linear combination of the two sides using coefficients $1-c_m$ and c_m, there is

$$\begin{aligned}&\textstyle\sum_{j=1}^{J} \left[\left(1-c_m\right) n_j p_j \mathbb{1}\left(q_j \geqslant c_h\right) + c_m \left(1-p_j\right) n_j \mathbb{1}\left(q_j < c_h\right)\right] \\ > \;&\textstyle\sum_{j=1}^{J} \left[\left(1-c_m\right) n_j p_j \mathbb{1}\left(p_j \geqslant c_m\right) + c_m \left(1-p_j\right) n_j \mathbb{1}\left(p_j < c_m\right)\right]\end{aligned}$$

This is clearly a contradiction since the right hand side achieves

$$\max_{\delta_j, j=1,\ldots,J} \sum_{j=1}^{J} \left[\left(1-c_m\right) n_j p_j \delta_j + c_m \left(1-p_j\right) n_j \left(1-\delta_j\right)\right].$$

Next we provide a concavity result, similar to the continuous limit population case, that are depicted in figure 18,

[Figure 18 about here.]

where adjacent machine TPR/FPR pairs are connected by a straight line to form a ROC "curve". The slopes of these connecting lines can be shown to be monotonically decreasing. First the slope of the line connecting the pairs between (T_{k+1}, F_{k+1}) and (T_k, F_k) can be written as, up to a constant, supposing p_k is ranked in decreasing order

$$s_k = \frac{\sum_{j=1}^{J} n_j p_j \mathbb{1}\left(p_j \geqslant p_k\right) - \sum_{j=1}^{J} n_j p_j \mathbb{1}\left(p_j \geqslant p_{k+1}\right)}{\sum_{j=1}^{J} n_j \left(1-p_j\right) \mathbb{1}\left(p_j \geqslant p_k\right) - \sum_{j=1}^{J} n_j \left(1-p_j\right) \mathbb{1}\left(p_j \geqslant p_{k+1}\right)} = \frac{p_k}{1-p_k}.$$

This is an increasing function of p_k, meaning that the slope decreases from the $(0,0)$ corner, where cutoff $c_m = p_k$ is close to 1, to the $(1,1)$ corner, where the cutoff $c_m = p_k$ decreases to close to 0. Hence the ROC curve represented by interpolating the $J+1$ points with straight lines is a concave curve.

In particular, the line segment between any given two points (T_{k+1}, F_{k+1}) and (T_k, F_k)

Electronic copy available at: https://ssrn.com/abstract=3508224

represents the FPR/TPR pairs of a collection of randomized decision rules:

$$(1-\lambda)\,\mathbb{1}\left(p_j \geqslant p_{k+1}\right) + \lambda \mathbb{1}\left(p_j \geqslant p_k\right),$$

where λ ranges from $(0,1)$. This corresponds to a randomized classification rule utilizing an independent uniform random variable $U_i \sim U(0,1)$, such that

$$\hat{y}_i = \begin{cases} 1 & p(x_i) \geqslant p_k \quad \text{and} \quad U_i \leqslant \lambda \\ 0 & p(x_i) < p_k \quad \text{and} \quad U_i < \lambda \\ 1 & p(x_i) \geqslant p_{k+1} \quad \text{and} \quad U_i \geqslant \lambda \\ 0 & p(x_i) \geqslant p_{k+1} \quad \text{and} \quad U_i < \lambda \end{cases}$$

In fact, any concave ROC curve can be dominated by a concavified ROC using linear extrapolation that corresponds to randomized decision rules. Consider two points (F_1, T_1) and (F_2, T_2) on the FPR/TPR plane that correspond to classification rules of $\hat{y} = \mathbb{1}(x \in R_1)$ and $\hat{y} = \mathbb{1}(x \in R_2)$. Let the new randomized decision rule be

$$\hat{y}_i = \mathbb{1}\left(U_i \geqslant \lambda\right)\mathbb{1}\left(X_i \in R_1\right) + \mathbb{1}\left(U_i < \lambda\right)\mathbb{1}\left(X_i \in R_2\right).$$

The FPR/TPR pair corresponding to this randomized classification rule is then easily computed to be a linear combination

$$\left((1-\lambda)F_1 + \lambda F_2, (1-\lambda)T_1 + \lambda T_2\right).$$

When λ ranges from $(0,1)$, this traces out the line segment between (F_1, T_1) and (F_2, T_2).

A.3 Proof of Asymptotic Joint Normal Distribution of FPR/TPR Pair

Before we show the joint normal distribution of FPR/TPR pair, we first show the asymptotic distribution of TPR and FPR of human decision makers. We then present the joint distribution.

A.3.1 Asymptotic Distribution of TPR

Denote p_0 as population probability of positive events, and p_1 and p_2 as TPR and FPR in the population. For briefness, we write TPR as $\widehat{A}$, i.e.,

$$\widehat{A} = \frac{\sum_{i=1}^{n} Y_i \widehat{Y}_i}{\sum_{i=1}^{n} Y_i}$$

Electronic copy available at: https://ssrn.com/abstract=3508224

and let

$$A = \frac{\mathbb{E}Y_i\widehat{Y}_i}{\mathbb{E}Y_i} = p_1.$$

We want to obtain the limiting distribution of $\widehat{A}$, of the form of

$$\sqrt{n}(\widehat{A} - A) \xrightarrow{d} N(0, \zeta^2)$$

where

$$\sqrt{n}(\widehat{A} - A) = \frac{\sqrt{n}}{n\widehat{p}} \sum_{i=1}^{n} (Y_i\widehat{Y}_i - \widehat{p}A).$$

By weak law of large number,

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{\mathbb{P}} p_0.$$

So, by Slutsky's theorem, it suffices to show that for a constant Ω

$$\frac{\sqrt{n}}{n} \sum_{i=1}^{n} (Y_i\widehat{Y}_i - \widehat{p}A) \xrightarrow{d} N(0, \Omega),$$

then

$$\sqrt{n}(\widehat{A} - A) \xrightarrow{d} N(0, \frac{\Omega}{p_0^2}).$$

Since

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

and

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} \widehat{p},$$

$$\frac{\sqrt{n}}{n} \sum_{i=1}^{n} (Y_i\widehat{Y}_i - \widehat{p}A) = \frac{\sqrt{n}}{n} \sum_{i=1}^{n} (Y_i\widehat{Y}_i - AY_i).$$

We write

$$\eta(Y_i, \widehat{Y}_i) = Y_i\widehat{Y}_i - AY_i,$$

$$\frac{\sqrt{n}}{n} \sum_{i=1}^{n} (Y_i\widehat{Y}_i - \widehat{p}A) = \frac{\sqrt{n}}{n} \sum_{i=1}^{n} \eta(Y_i, \widehat{Y}_i).$$

By the De Moivre-Laplace central limit theorem,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta(Y_i, \widehat{Y}_i) \xrightarrow{d} N(0, \Omega),$$

Electronic copy available at: https://ssrn.com/abstract=3508224

where

$$\Omega = Var(\eta(Y_i, \widehat{Y}_i)),$$

where

$$\begin{aligned}\mathbb{E}\eta &= -Ap_0(1-p_1) + (1-A)p_0p_1 \\ &= -Ap_0 + p_0p_1\end{aligned}$$

$$\begin{aligned}\mathbb{E}\eta^2 &= A^2p_0(1-p_1) + (1-A)^2p_0p_1 \\ &= A^2p_0 - 2Ap_0p_1 + p_0p_1\end{aligned}$$

so

$$\begin{aligned}Var(\eta(Y_i, \widehat{Y}_i)) &= \mathbb{E}\eta^2 - (\mathbb{E}\eta)^2 \\ &= p_0p_1 - p_0p_1^2\end{aligned}$$

then we get

$$\sqrt{n}(\widehat{A} - A) \xrightarrow{d} N(0, \frac{p_1 - p_1^2}{p_0})$$

A.3.2 Asymptotic Normality of FPR

For briefness, we write FPR as $\widehat{B}$, i.e.,

$$\widehat{B} = \frac{\sum_{i=1}^n (1-Y_i)\widehat{Y}_i}{\sum_{i=1}^n (1-Y_i)}$$

and let

$$B = \frac{\mathbb{E}(1-Y_i)\widehat{Y}_i}{\mathbb{E}(1-Y_i)} = p_2.$$

Follow the definitions above, we have $Q = 1 - Y$, then $Q \sim B(1, 1-p_0)$, and

$$\widehat{B} = \frac{\sum_{i=1}^n Q_i\widehat{Y}_i}{\sum_{i=1}^n Q_i},$$

$$B = \frac{\mathbb{E}Q_i\widehat{Y}_i}{\mathbb{E}Q_i} = p_2.$$

Notice that $\widehat{B}$ and B have the same form as $\widehat{A}$ and A in section above, so we could obtain the limiting distribution of $\widehat{B}$ in a similar way, i.e.,

Electronic copy available at: https://ssrn.com/abstract=3508224

$$\sqrt{n}(\widehat{B}-B) \xrightarrow{d} N(0, \frac{p_2-p_2^2}{1-p_0})$$

A.3.3 Asymptotic Joint Normality of FPR/TPR Pair

We have proved the asymptotic normality of both FPR and TPR in sections above, and we want to obtain the asymptotic normality of (FPR, TPR).

Let $p_0 = EY_i$, $p_1 = \frac{EY_i\hat{Y}_i}{EY_i}$, $p_2 = \frac{E(1-Y_i)\hat{Y}_i}{E(1-Y_i)}$. We want to obtain the asymptotic normality of (TPR, FPR). For briefness, we write this pair as $\hat{A}$ and let $A = (p_1, p_2)$. We want to obtain the limiting distribution of $\hat{A}$, of the form of

$$\sqrt{n}\left(\hat{A}-A\right) \xrightarrow{d} N\left(0,\Sigma\right).$$

We have $\hat{p} = \frac{1}{n}\sum_{i=1}^n Y_i \xrightarrow{\mathbb{P}} p_0$ and

$$\begin{aligned}\sqrt{n}\left(\hat{A}-A\right) =& \sqrt{n}\left(\frac{\sum_{i=1}^n Y_i\hat{Y}_i}{\sum_{i=1}^n Y_i} - p_1, \frac{\sum_{i=1}^n (1-Y_i)\hat{Y}_i}{\sum_{i=1}^n (1-Y_i)} - p_2\right)\\ =& \sqrt{n}\left(\frac{\sum_{i=1}^n \left[Y_i\hat{Y}_i - Y_i p_1\right]}{n\hat{p}}, \frac{\sum_{i=1}^n \left[(1-Y_i)\left(\hat{Y}_i - p_2\right)\right]}{n(1-\hat{p})}\right).\end{aligned}$$

Since $(\hat{p}, 1-\hat{p}) \xrightarrow{\mathbb{P}} (p_0, 1-p_0)$, by the multivariate Slutsky's theorem, it suffices to show that for a matrix Ω

$$\frac{\sqrt{n}}{n}\left(\sum_{i=1}^n Y_i\left(\hat{Y}_i - p_1\right), \sum_{i=1}^n (1-Y_i)\left(\hat{Y}_i - p_2\right)\right) \xrightarrow{d} N\left(0,\Omega\right).$$

The left side is $\frac{\sqrt{n}}{n}\sum_{i=1}^n \left(Y_i\left(\hat{Y}_i - p_1\right), (1-Y_i)\left(\hat{Y}_i - p_2\right)\right)$. Write the term in summation as (η_i, ξ_i). By the multivariate Central Limit Theorem

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (\eta_i, \xi_i) \xrightarrow{\mathbb{P}} N\left(0,\Sigma\right),$$

where $\Sigma = Cov\left((\eta_i, \xi_i)\right)$. We can calculate that

$$\Sigma = \begin{Bmatrix} p_0p_1 - p_0p_1^2 & 0 \\ 0 & p_0p_2 - p_0p_2^2 \end{Bmatrix},$$

Electronic copy available at: https://ssrn.com/abstract=3508224

because $Y_i(1-Y_i)\left(\hat{Y}_i - p_1\right)\left(\hat{Y}_i - p_2\right) \equiv 0$. So we get

$$\sqrt{n}\left(\hat{A} - A\right) \xrightarrow{d} N\left(0, \begin{Bmatrix} \frac{p_1 - p_1^2}{p_0} & 0 \\ 0 & \frac{p_2 - p_2^2}{p_0} \end{Bmatrix}\right). \tag{13}$$

This explain why the two-dimensional confidence interval is a ellipse (not a rotated ellipse).

A.4 Machine's Algorithm: The Random Forest

Consider a dataset $\mathbf{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_k, y_k)\}$, where each input vector $\mathbf{x} \in \mathbb{R}^d$ has d features. A random forest algorithm with N estimators and M max features per node, i.e. for each node in a tree, no more then M features will be considered to obtain the best split, works as follows:

1. For $i = 1$ to N:
   (a) Draw a bootstrap sample D_i from the training data $\mathbf{D}$; i.e. D_i has the same sample size as $\mathbf{D}$ but is drawn with replacement.
   (b) Grow an unpruned tree T_i using D_i by repeating the following steps for each node of the tree, until the leaves are pure or the minimum number of samples required to split is reached:
      i. Randomly choose M features from d features of input vector.
      ii. Pick the best one feature to split among M features using Gini impurity criterion.
      iii. Split the node into two sub-nodes using the best feature.
2. Obtain the random forest output $\{T_i\}_{i=1}^N$.

Given a new input $\mathbf{x}$, the random forest makes a prediction by aggregating the results of N trees:

- Regression: $\widehat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^N T_i(\mathbf{x})$.
- Classification: $\widehat{C}(\mathbf{x})$ = majority vote of $\{T_i(\mathbf{x})\}_{i=1}^N$. The probability of class $\widehat{p}(\widehat{C}(\mathbf{x}))$ is computed as the mean of class probability of each tree in the forest. The class probability of a single tree is the fraction of samples of the same class in a leaf of which $\mathbf{x}$ falls into.

Electronic copy available at: https://ssrn.com/abstract=3508224

A.5 Overview of Algorithm

The overview algorithm of approach based on asymptotic confidence set with random ROC sampling (Section 4.1) is as follows:

1. **Training RF Model and Computing the Machine's ROC Curve.** Split the training subset of observations for each doctor into two parts: the first part is aggregated and used to grow a random forest model as described in section A.4; the features and label from the second part are used as inputs to make prediction by the RF model, and to tabulate the machine's ROC curve.

2. **Classifying Capable and Incapable Doctors.** Classify each doctor into the capable or incapable group using the methods in section 2.4 with his/her training set as historical records.

3. **Generating Thresholds Interval of Incapable Doctors.** For each doctor d_i in incapable group, obtain the threshold interval endpoints $c_{d_i,1}$ and $c_{d_i,2}$. As described in section 2.4, we first obtain the point $P(x_p, y_p)$ that represents the lowest FPR x_p and highest TPR y_p of doctor's confidence set. Denote the decision corresponding to $(x_p, TPR|_{x_p})$ by $c_{d_i,2}$, and the decision threshold corresponding to $(FPR|_{y_p}, y_p)$ on machine's ROC curve by $c_{d_i,1}$. The machine's performance with any threshold $c_{d_i} \in [c_{d_i,1}, c_{d_i,2}]$ would be better than this doctor's.

4. **Making Prediction and Decision.** Make prediction and decision on the testing set. For capable doctors, their decisions are not replaced by machines, and are thus kept as final decisions.

   For incapable doctors, generate machine's decisions with specific threshold c_{d_i} with the following algorithm. For each doctor d_i and each corresponding threshold c_{d_i}, random forest model makes prediction $\widehat{p}(x)$ on each case x the in testing set, and generates decision using the specific threshold c_{d_i}: $\widehat{D}(x) = \mathbb{1}(\widehat{p}(x) > c_{d_i})$.

5. **Evaluating the Overall Performance.** Combine the decisions of capable doctors and machine algorithm on the testing set to obtain the overall FPR and TPR.

   First, we initialize a single FPR/TPR pair, where the threshold of each incapable doctor d_i is set to $c_{d_i,1}$, which corresponds to point of $(x_p, TPR|_{x_p})$ on machine's ROC curve.

   Second, we trace out the machine algorithm's performance for threshold intervals of incapable doctors by uniformly sampling N FPR/TPR pairs using following algorithms.

Electronic copy available at: https://ssrn.com/abstract=3508224

(a) For each incapable doctor d_i, set the step of threshold $c_{d_i,step} = (c_{d_i,2} - c_{d_i,1})/N$

(b) For each $j = 1$ to N:

Obtain a single FPR/TPR pair result T_j where the threshold of each incapable doctor d_i is set to $c_{d_i,1} + c_{d_i,step} * j$

(c) Collect the uniformly sampled FPR/TPR pairs $\{T_j\}_{j=1}^{N}$.

References

Acemoglu, Daron and Pascual Restrepo, "Secular stagnation? The effect of aging on economic growth in the age of automation," *American Economic Review*, 2017, *107* (5), 174–79.

_ and _ , "Low-skill and high-skill automation," *Journal of Human Capital*, 2018, *12* (2), 204–232.

_ and _ , "The race between man and machine: Implications of technology for growth, factor shares, and employment," *American Economic Review*, 2018, *108* (6), 1488–1542.

Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preoiuc-Pietro, and Vasileios Lampos, "Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective," *PeerJ Computer Science*, 2016, *2*, e93.

Berk, Richard, "An impact assessment of machine learning risk forecasts on parole board decisions and recidivism," *Journal of Experimental Criminology*, 2017, *13* (2), 193–216.

Breiman, Leo, "Random forests," *Machine learning*, 2001, *45* (1), 5–32.

Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock, "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?," in "AEA Papers and Proceedings," Vol. 108 2018, pp. 43–47.

Castro, Victor M, Dmitriy Dligach, Sean Finan, Sheng Yu, Anil Can, Muhammad Abd-El-Barr, Vivian Gainer, Nancy A Shadick, Shawn Murphy, Tianxi Cai et al., "Large-scale identification of patients with cerebral aneurysms using natural language processing," *Neurology*, 2017, *88* (2), 164–168.

Chandra, Amitabh and Douglas O Staiger, "Expertise, underuse, and overuse in healthcare," *Working paper*, 2011.

Chernozhukov, Victor and Han Hong, "An MCMC approach to classical estimation," *Journal of Econometrics*, 2003, *115* (2), 293–346.

Electronic copy available at: https://ssrn.com/abstract=3508224

Currie, Janet and WB MacLeod, "Diagnosing expertise: Human capital, decision making, and performance among physicians," *Journal of labor economics*, 2017, *35* (1), 1–43.

Esteva, Andre, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, 2017, *542* (7639), 115.

Everett, Michael, Yu Fan Chen, and Jonathan P How, "Motion planning among dynamic, decision-making agents with deep reinforcement learning," in "2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)" IEEE 2018, pp. 3052–3059.

Feng, Kai, Han Hong, Ke Tang, and Jingyuan Wang, "Decision making with machine learning and ROC curves," 2019. Working paper.

Gulshan, Varun, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, 2016, *316* (22), 2402–2410.

Kermany, Daniel S, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, 2018, *172* (5), 1122–1131.

Keynes, John Maynard, "Economic possibilities for our grandchildren," in "Essays in persuasion," Springer, 2010, pp. 321–332.

Kim, Jae-Young, "Limited information likelihood and Bayesian analysis," *Journal of Econometrics*, 2002, *107* (1-2), 175–193.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The Quarterly Journal of Economics*, 2018, *133* (1), 237–293.

Liang, Huiying, Brian Y. Tsui, Hao Ni, Carolina C. S. Valentim, Sally L. Baxter, Guangjian Liu, Wenjia Cai, Daniel S. Kermany, Xin Sun, Jiancong Chen et al., "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence," *Nature Medicine*, 2019, *25* (3), 433–438.

Electronic copy available at: https://ssrn.com/abstract=3508224

Liu, Yun, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado et al., "Detecting cancer metastases on gigapixel pathology images," *arXiv preprint arXiv:1703.02442*, 2017.

Nieto, Yuri, Vicente García-Díaz, Carlos Montenegro, and Rubén González Crespo, "Supporting academic decision making at higher educational institutions using machine learning-based algorithms," *Soft Computing*, 2019, *23* (12), 4145–4153.

Onu, Charles C, Innocent Udeogu, Eyenimi Ndiomu, Urbain Kengni, Doina Precup, Guilherme M Sant'Anna, Edward Alikor, and Peace Opara, "Ubenwa: Cry-based diagnosis of birth asphyxia," *arXiv preprint arXiv:1711.06405*, 2017.

Rajpurkar, Pranav, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," *arXiv preprint arXiv:1707.01836*, 2017.

Rasch, R., A. Kott, and K.D. Forbus, "Incorporating AI into military decision making: an experiment," *IEEE Intelligent Systems*, 2003, *18* (4), 18–26.

Razzaki, Salman, Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, Daniel Mullarkey, Davinder Sangar, Michael Taliercio, Mobasher Butt, Azeem Majeed, Arnold DoRosario, Megan Mahoney, and Saurabh Johri, "A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis," *arXiv:1806.10698 [cs, stat]*, 2018. arXiv: 1806.10698.

Reis, MAM, NRS Ortega, and Paulo Sérgio Panse Silveira, "Fuzzy expert system in the prediction of neonatal resuscitation," *Brazilian Journal of Medical and Biological Research*, 2004, *37* (5), 755–764.

Rudin, C., D. Waltz, R. N. Anderson, A. Boulanger, A. Salleb-Aouissi, M. Chow, H. Dutta, P. N. Gross, B. Huang, S. Ierome, D. F. Isaac, A. Kressner, R. J. Passonneau, A. Radeva, and L. Wu, "Machine Learning for the New York City Power Grid," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, *34* (2), 328–345.

Wang, JunPing, YunChuan Sun, WenSheng Zhang, Ian Thomas, ShiHui Duan, and YouKang Shi, "Large-scale online multitask learning and decision making for flexible manufacturing," *IEEE Transactions on Industrial Informatics*, 2016, *12* (6), 2139–2147.

Electronic copy available at: https://ssrn.com/abstract=3508224

Table 1: Replaced Doctors and Geographic Characteristics

Panel A: Bayesian Approach						
	(1) OLS	(2) OLS	(3) OLS	(4) Logit	(5) Logit	(6) Logit
gdp_all_province	-0.0971		-0.0773	-0.393		-0.317
	(-0.74)		(-0.59)	(-0.74)		(-0.59)
township=1		0.103	0.101		0.417	0.411
		(2.44)	(2.40)		(2.43)	(2.39)
Constant	0.490	0.402	0.427	-0.0373	-0.399	-0.296
	(10.59)	(12.63)	(8.02)	(-0.20)	(-3.05)	(-1.37)
Observations	563	563	563	563	563	563
R^2	0.001	0.011	0.011			
Panel B: Frequentist Approach						
	(1) OLS	(2) OLS	(3) OLS	(4) Logit	(5) Logit	(6) Logit
gdp_all_province	-0.184		-0.164	-0.752		-0.678
	(-1.40)		(-1.25)	(-1.40)		(-1.25)
township=1		0.108	0.105		0.439	0.427
		(2.56)	(2.48)		(2.55)	(2.47)
Constant	0.513	0.393	0.447	0.0533	-0.433	-0.214
	(11.09)	(12.39)	(8.41)	(0.28)	(-3.30)	(-0.98)
Observations	563	563	563	563	563	563
R^2	0.004	0.012	0.014			

t statistics in parentheses

Electronic copy available at: https://ssrn.com/abstract=3508224

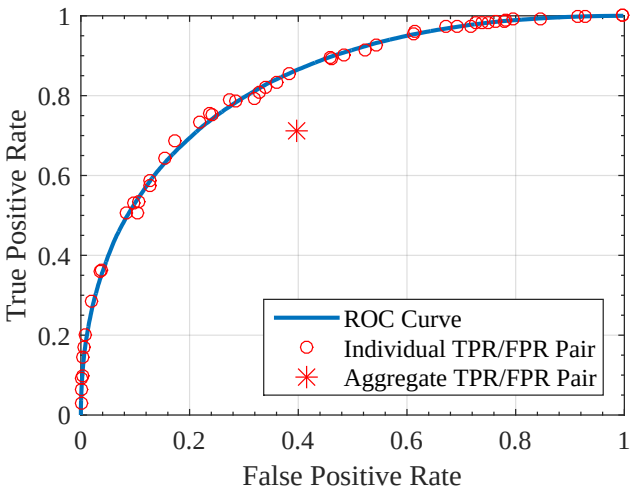


Figure 1: Individual and Aggregate TPR/FPR Pairs: Perspective of Jensen's Inequality

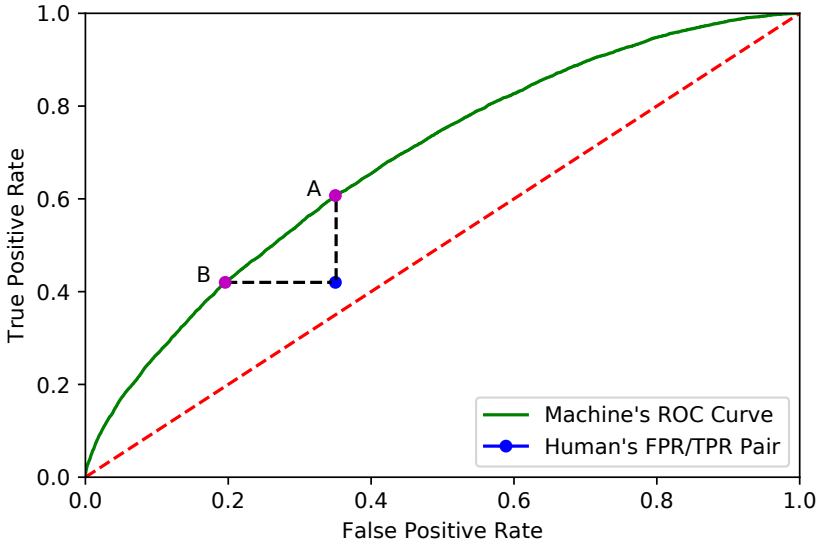


Figure 2: Human's FPR/TPR Pair and Machine's ROC Curve

Electronic copy available at: https://ssrn.com/abstract=3508224

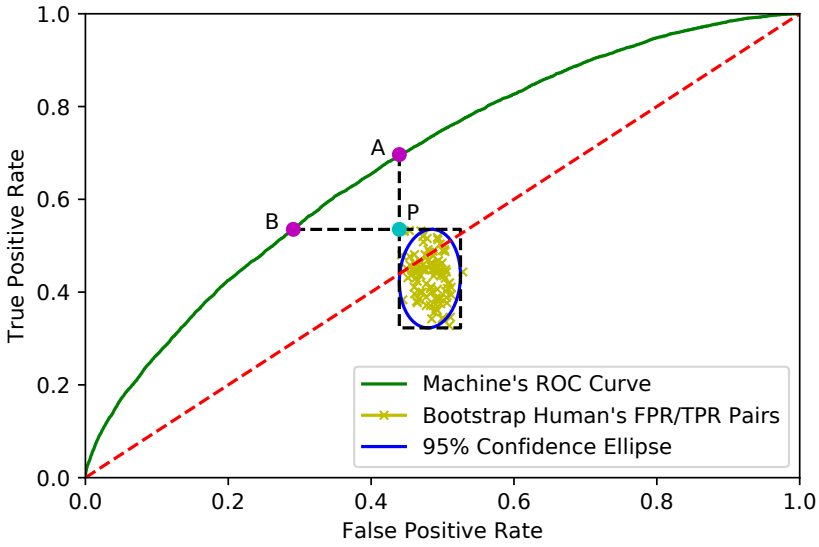


Figure 3: Sketch Map of Our Method for A Single Decision Maker

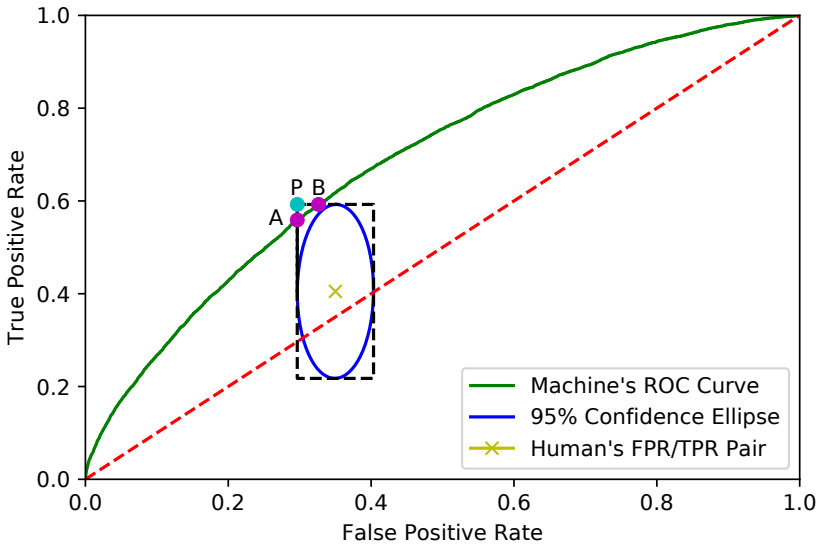


Figure 4: Case 2

Electronic copy available at: https://ssrn.com/abstract=3508224

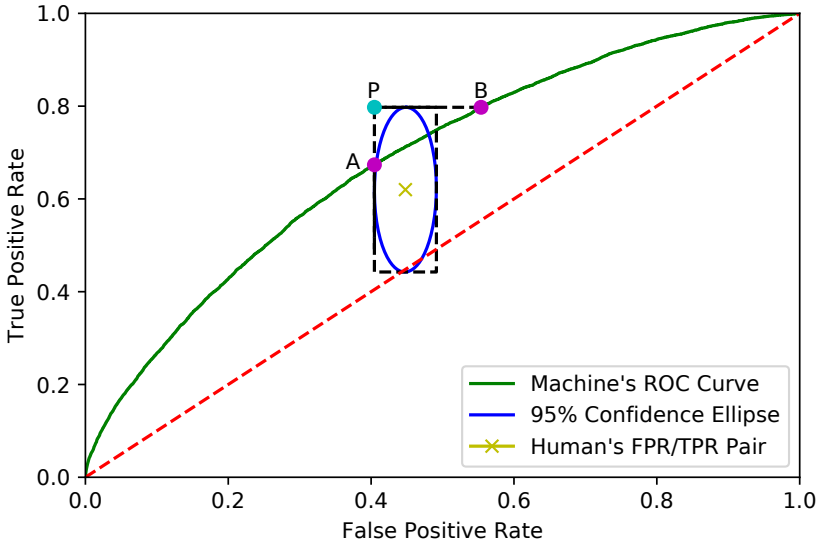


Figure 5: Case 3

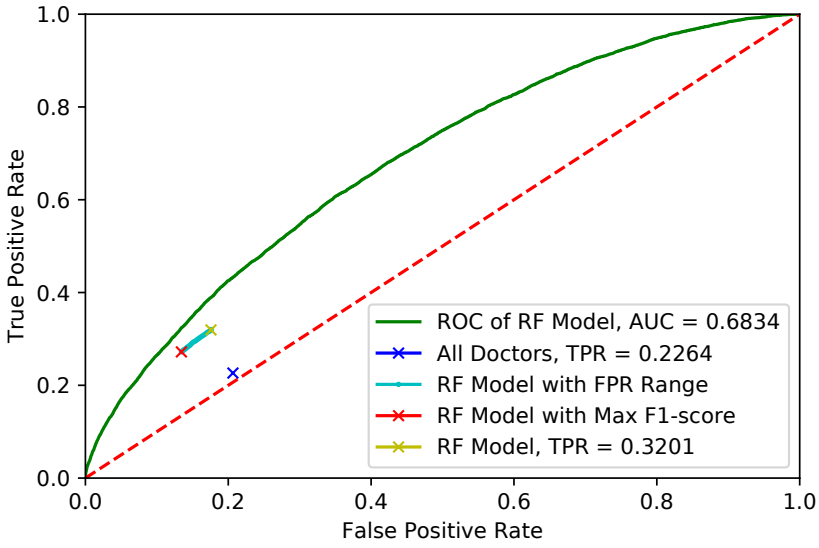


Figure 6: Performance of RF Model and Doctors (Doctor's Diagnosis >= 300)

Electronic copy available at: https://ssrn.com/abstract=3508224

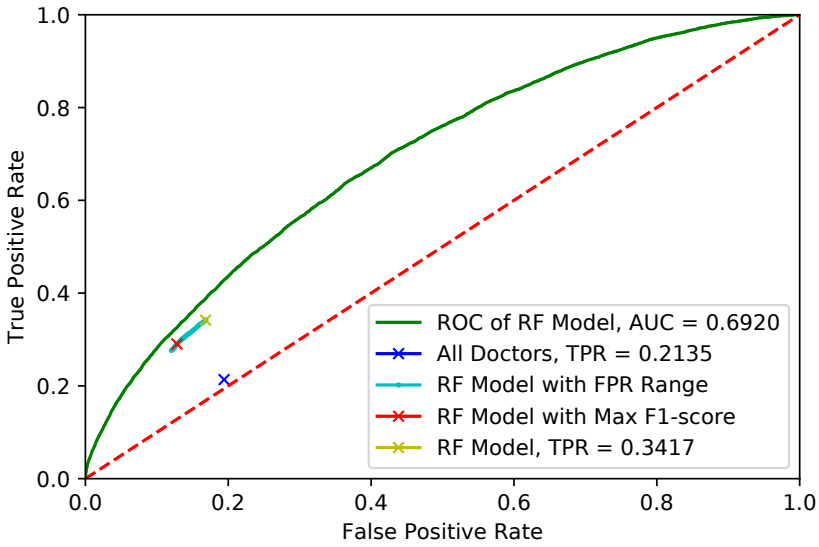


Figure 7: Performance of RF Model and Doctors (Doctor's Diagnosis >= 500)

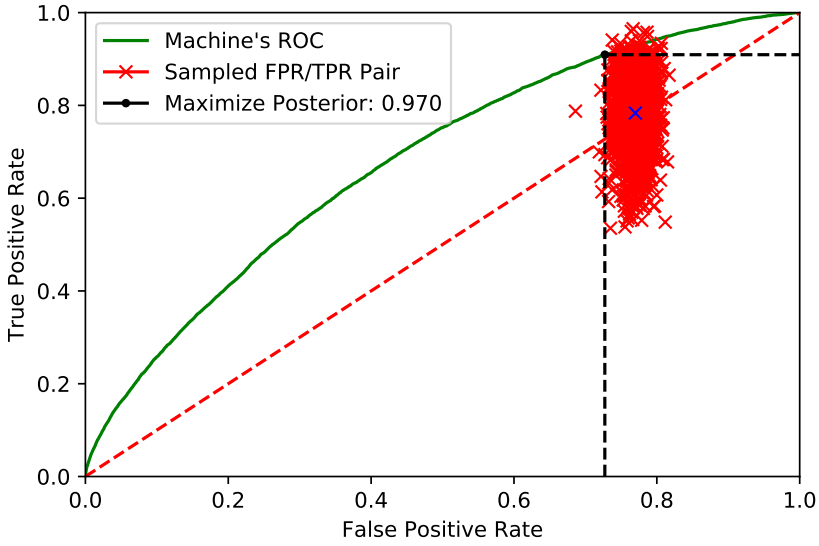


Figure 8: Bayesian Case 1

Electronic copy available at: https://ssrn.com/abstract=3508224

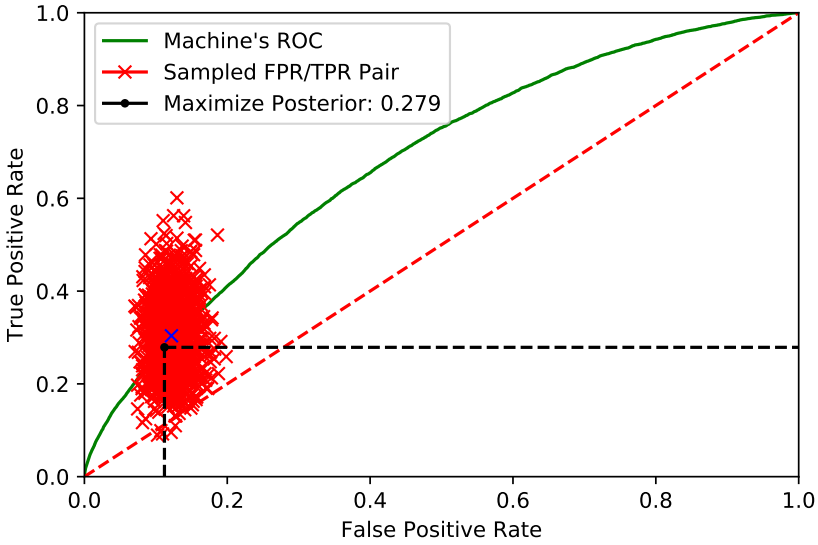
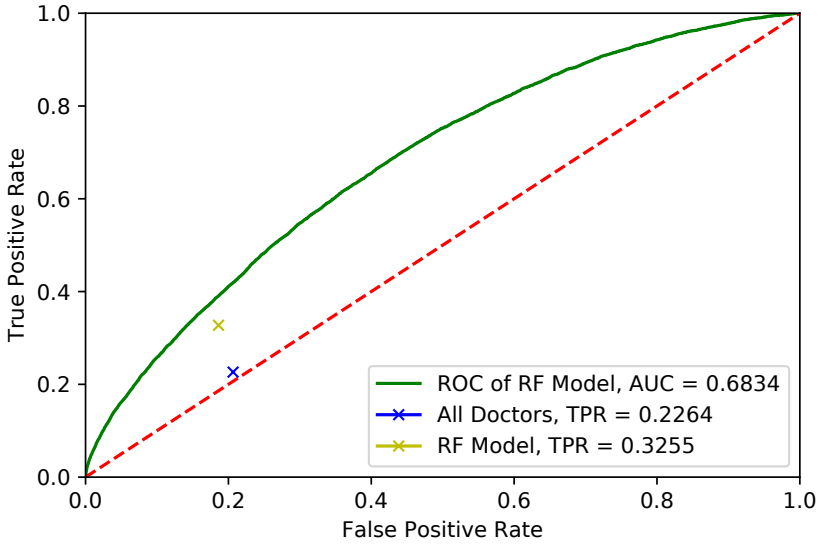
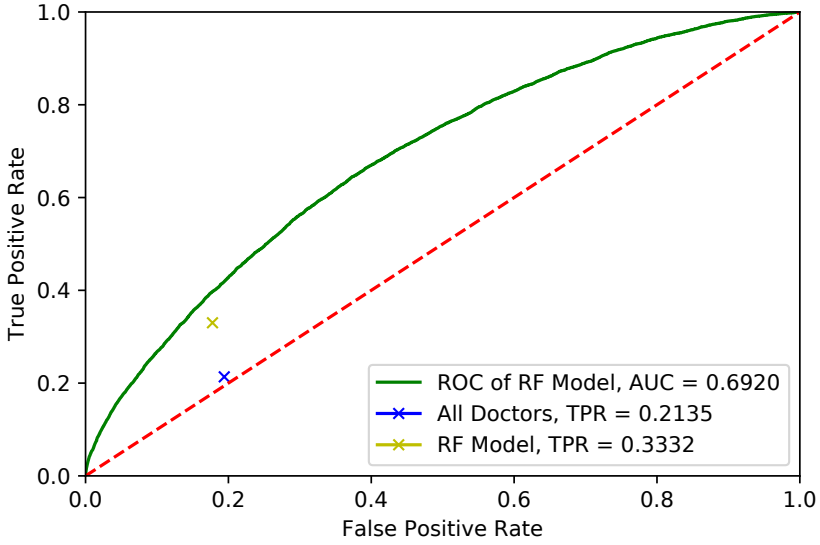


Figure 9: Bayesian Case 2

Electronic copy available at: https://ssrn.com/abstract=3508224



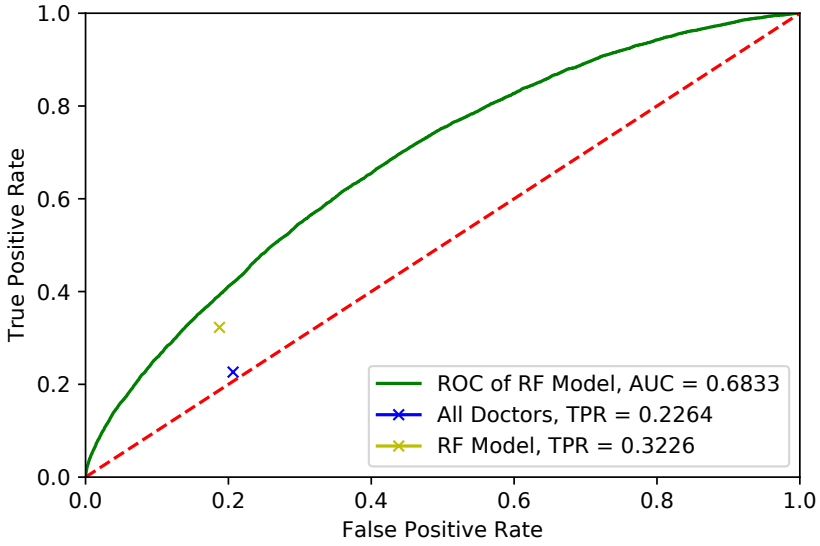
(a) Doctor's Diagnosis >= 300



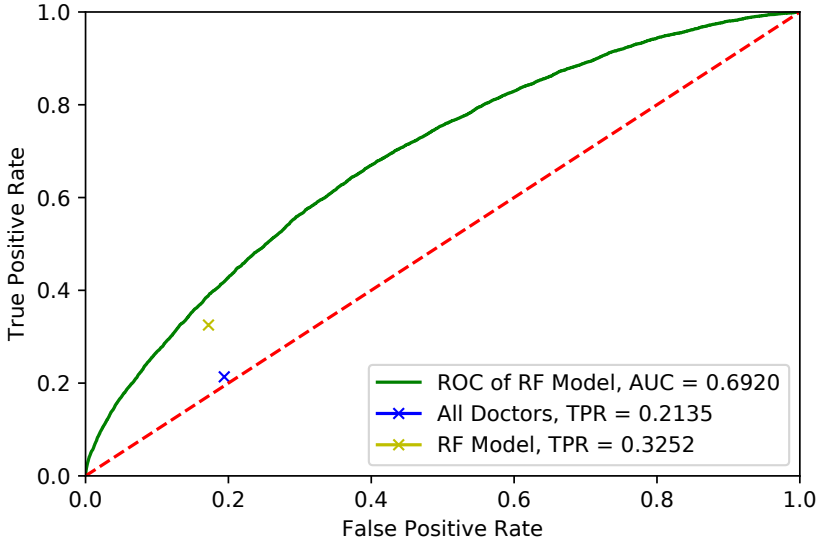
(b) Doctor's Diagnosis >= 500

Figure 10: Result of Bayesian Approach

Electronic copy available at: https://ssrn.com/abstract=3508224



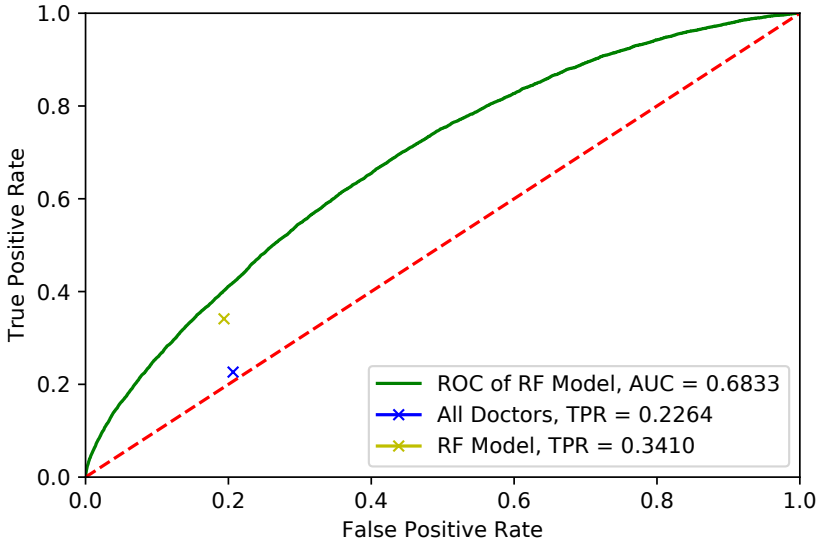
(a) Doctor's Diagnosis >= 300



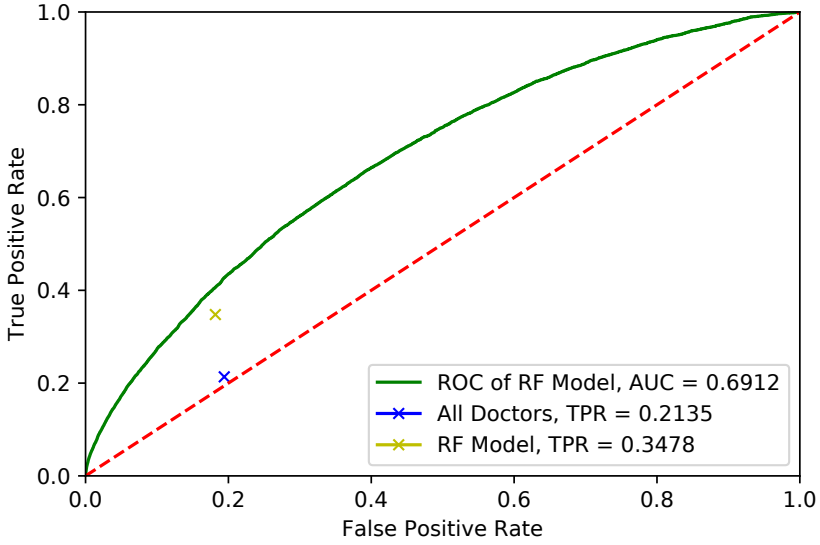
(b) Doctor's Diagnosis >= 500

Figure 11: Result of Frequentist Approach

Electronic copy available at: https://ssrn.com/abstract=3508224



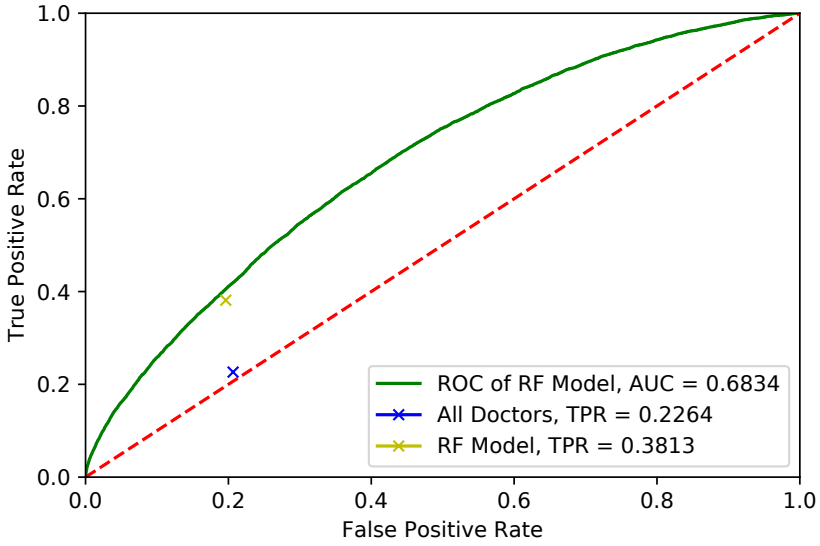
(a) Doctor's Diagnosis >= 300



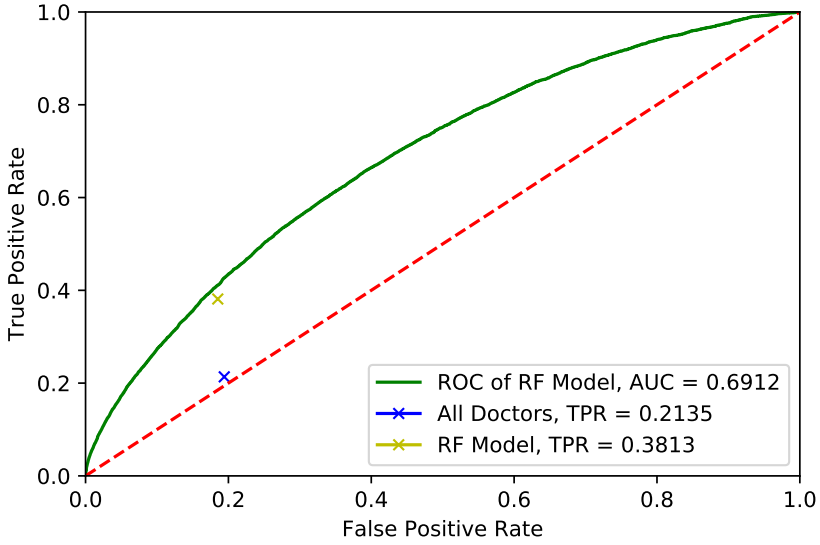
(b) Doctor's Diagnosis >= 500

Figure 12: Result of Frequentist Approach 2

Electronic copy available at: https://ssrn.com/abstract=3508224



(a) Doctor's Diagnosis >= 300



(b) Doctor's Diagnosis >= 500

Figure 13: Result of Frequentist Approach 3

Electronic copy available at: https://ssrn.com/abstract=3508224

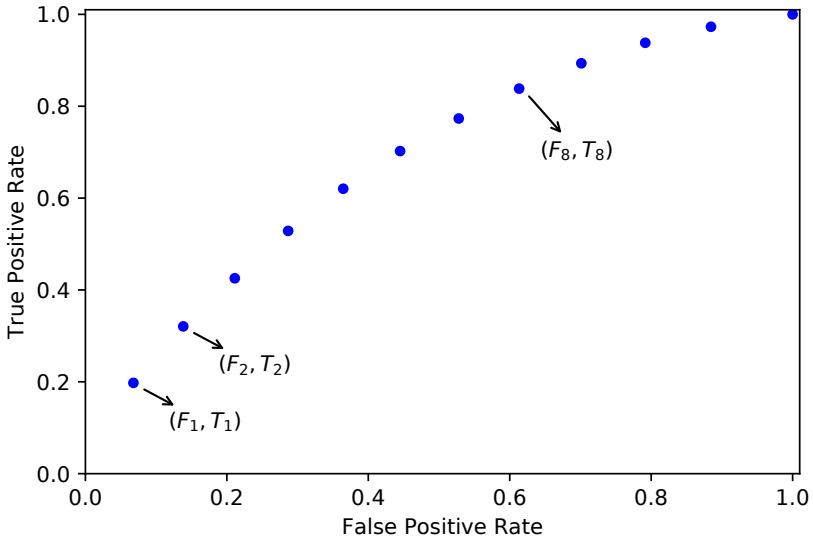


Figure 14: Finite Sample ROC points

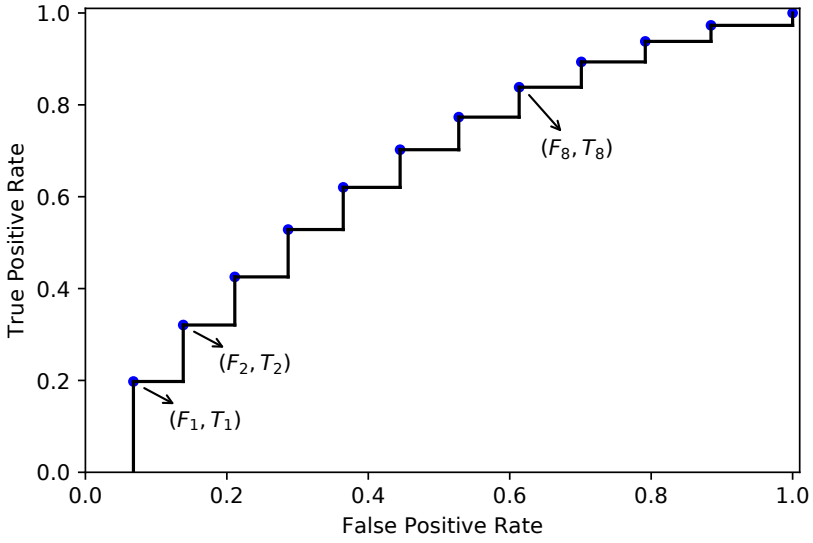


Figure 15: Finite Sample ROC points

Electronic copy available at: https://ssrn.com/abstract=3508224

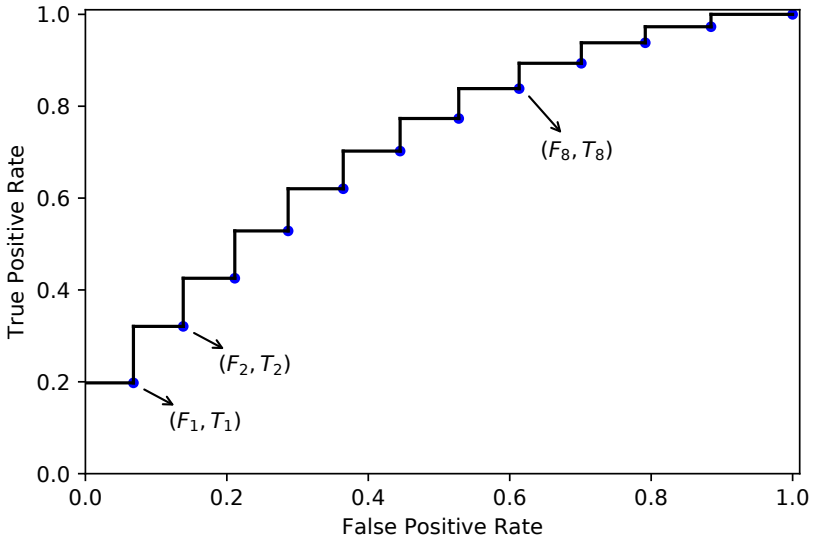


Figure 16: Finite Sample ROC points

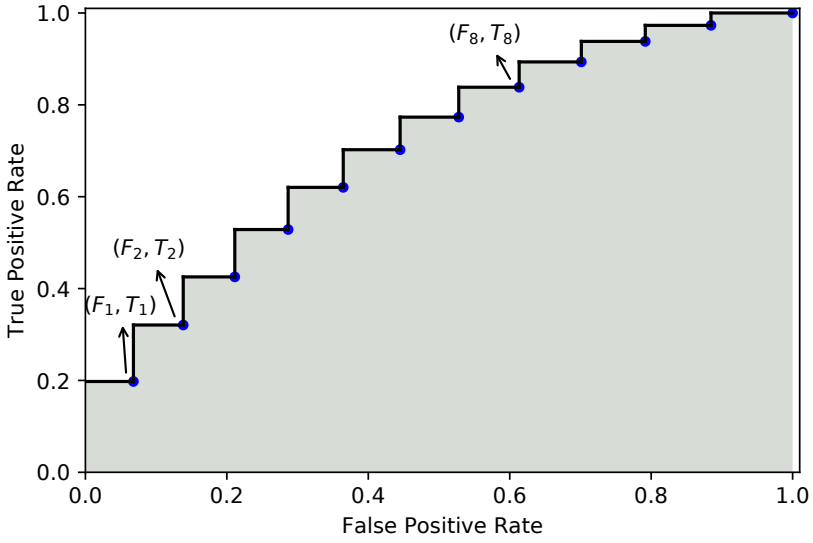


Figure 17: Finite Sample ROC points

Electronic copy available at: https://ssrn.com/abstract=3508224

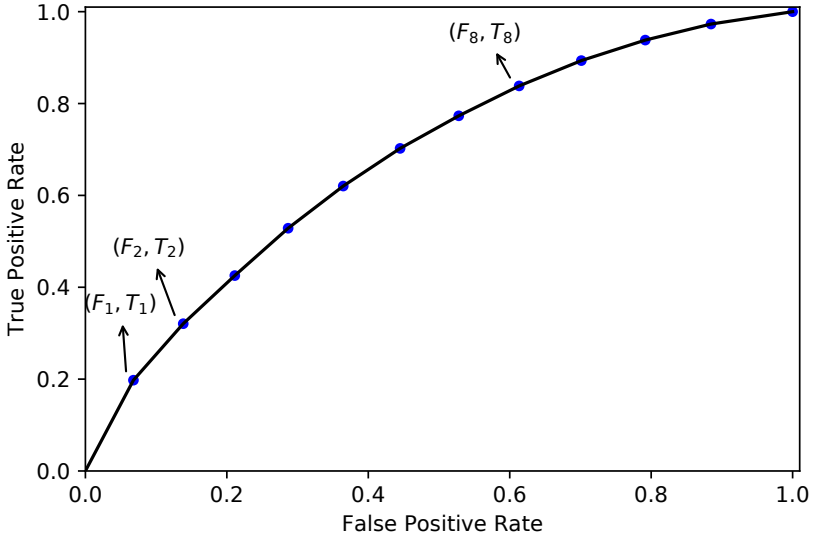


Figure 18: Finite Sample ROC points

Electronic copy available at: https://ssrn.com/abstract=3508224