Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Statistical Tests for Replacing Human Decision Makers with Algorithms

Kai Feng; , Han Hong; , Ke Tang; , Jingyuan Wang

To cite this article:
Kai Feng; , Han Hong; , Ke Tang; , Jingyuan Wang (2025) Statistical Tests for Replacing Human Decision Makers with Algorithms. Management Science

Published online in Articles in Advance 05 Mar 2025

. https://doi.org/10.1287/mnsc.2023.01845

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–26
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Statistical Tests for Replacing Human Decision Makers with Algorithms

Kai Feng,[a] Han Hong,[b] Ke Tang,[a] Jingyuan Wang[c,*]

[a] Institute of Economics, School of Social Sciences, Tsinghua University, Beijing 100084, China; [b] Department of Economics, Stanford University, Stanford, California 94305; [c] School of Economics and Management, Beihang University, Beijing 100191, China
*Corresponding author
Contact: fengk22@mails.tsinghua.edu.cn (KF); doubleh@stanford.edu (HH); ketang@tsinghua.edu.cn (KT); jywang@buaa.edu.cn, https://orcid.org/0000-0003-0651-1592 (JW)

Received: June 27, 2023
Revised: September 10, 2024
Accepted: December 15, 2024
Published Online in Articles in Advance: March 5, 2025

https://doi.org/10.1287/mnsc.2023.01845

Abstract. This paper proposes a statistical framework of using artificial intelligence to improve human decision making. The performance of each human decision maker is benchmarked against that of machine predictions. We replace the diagnoses made by a subset of the decision makers with the recommendation from the machine learning algorithm. We apply both a heuristic frequentist approach and a Bayesian posterior loss function approach to abnormal birth detection using a nationwide data set of doctor diagnoses from prepregnancy checkups of reproductive-age couples and pregnancy outcomes. We find that our algorithm on a test data set results in a higher overall true positive rate and a lower false positive rate than the diagnoses made by doctors only.

History: Accepted by Yan Chen, behavioral economics and decision analysis.
Funding: H. Hong's work was supported by the National Science Foundation [Grant SES 1658950]. K. Tang's work was supported by the National Natural Science Foundation of China [Grants 72192802, 72342008]. J. Wang's work was partially supported by the National Natural Science Foundation of China [Grants 72222022, 72171013, 72242101].
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2023.01845.

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

1. Introduction

In the current era of machine learning, artificial intelligence (AI) algorithms have emerged as valuable tools that can learn significant features from data and potentially facilitate decision making in various disciplines. In medical research, Esteva et al. (2017) and Kermany et al. (2018) utilize a deep neural network to automate diagnosis based on medical images. Berg et al. (2020) and Sadhwani et al. (2021) leverage machine learning algorithms to predict the default probability for bank lending decisions. Mullainathan and Spiess (2017) and Athey and Imbens (2019) apply machine learning models to broad optimal policy assignment problems.

A considerable number of papers compare the performance of algorithms with that of the representative human decision makers. To mention a few, Esteva et al. (2017) claim that their algorithm outperforms the aggregate dermatologist on some skin cancer classification tasks; Kermany et al. (2018) target macular degeneration diagnosis and conclude that the machine algorithm outperforms some retinologists; Rajpurkar et al. (2017) highlight the performance gain of a deep convolutional neural network over an aggregate cardiologist for processing electrocardiogram sequences; Kleinberg et al. (2018) train a gradient-boosted decision tree model to predict crime risk during pretrial bail and suggest that substituting human judges with the algorithm could result in large welfare gains.

The findings of the aforementioned literature rely mostly on the observation that the pair of false positive rate (FPR) and true positive rate (TPR) point lies strictly below the receiver operating characteristic (ROC) curve generated by the machine learning algorithm, implying that algorithms can achieve a higher TPR for a given FPR or a lower FPR for a given TPR. In binary classification/decision making, FPR is the number of wrongly classified negative events divided by the number of actual negative events, and TPR is the number of correctly classified positive events divided by the number of total positive events. FPR/TPR are synonymous with size and power in classic hypothesis testing.

We first caution against such interpretations without a deeper understanding of human decision-making processes. The literature findings can be rationalized not only by the superior information quality of algorithms, but also by the incentive heterogeneity, that is, the differential preferences for taking the same action by multiple human decision makers, some of whom can be no

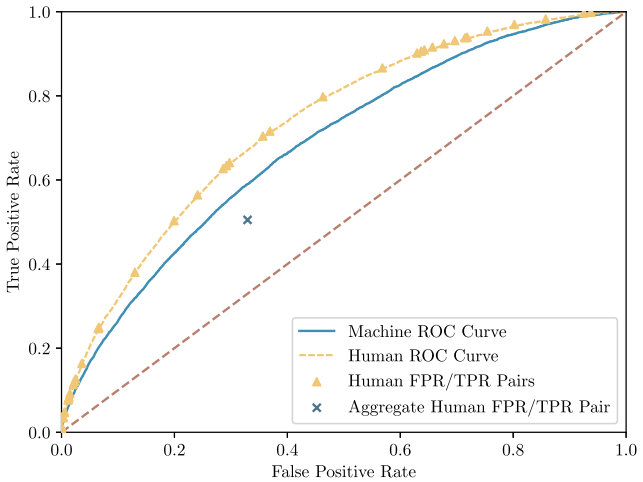
less accurate than algorithms in processing information from observational data. Machine algorithms provide information, but decision making is a combination of incentives and information. Using machine algorithms to assist with decision making necessitates the modeling of incentives. Second, the FPR/TPR pairs of human decision makers are typically imprecisely measured, especially when the number of cases for each decision maker is not large. The data in our empirical analysis shows not only substantial heterogeneity between individual doctors, but also that the randomness in measuring the quality of decision makers plays a key role when the performance of decision makers is compared with that of machine algorithms.

To illustrate the first issue of concern, consider Figure 1, in which a collection of FPR/TPR pairs for human decision makers, denoted as $j = 1, \ldots, J$, all lie approximately on an ROC curve generated by their employed decision rules $\hat{Y}_{i,j} = \mathbb{1}(p(X_{i,j}, U_{i,j}) > c_j)$ with different individual cutoff threshold c_j for decision maker j. In the decision rule, $X_{i,j}$ are observed features used in the machine learning algorithm; $U_{i,j}$ are private information only observable to the human decision makers, and $p(x,u) = \mathbb{P}(Y = 1 \mid x,u)$ is a propensity score function. Also drawn is the machine ROC curve based on classification rules $\hat{Y}_i = \mathbb{1}(p(X_i) > c)$, which collects the corresponding FPR/TPR pairs by varying the cutoff threshold c, and where $p(x) = \mathbb{P}(Y = 1 \mid x)$ is the correctly specified propensity score function using observable features x discovered by the machine learning algorithm. However, after averaging over all decision makers, the aggregate human FPR/TPR point lies visibly below the machine ROC curve.

Our paper makes several contributions. First, we offer a conceptual framework of information and incentives for interpreting comparisons between the machine algorithms and human decision makers. Second, we provide a model to identify human decision maker candidates for replacement by machine algorithms and to generate machine algorithm–based decisions that incorporate human decision maker preferences. The replacement model accounts for the sampling uncertainty of individual decision makers. Third, an empirical application to abnormal birth detection demonstrates that our proposed replacement algorithm significantly improves the overall performance of risky pregnancy detection. Fourth, we conduct an extensive synthetic data analysis that serves as guidance for evaluating empirical procedures incorporating complementarity in the presence of information asymmetry between human decision makers and machine algorithms.

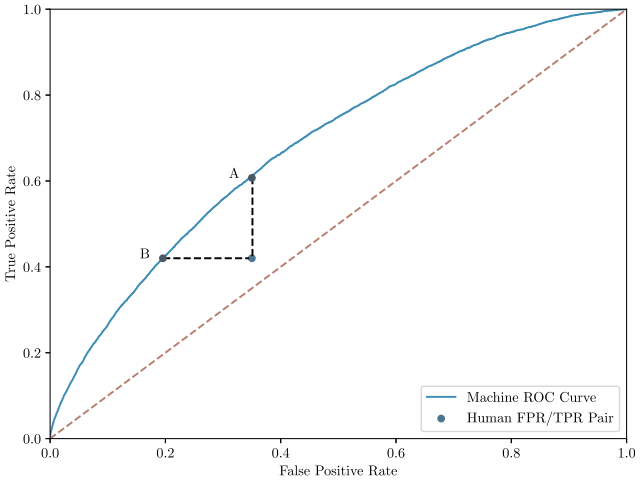
In this paper, in order to compare machines with human decision makers, we restrict incentive heterogeneity and focus on information processing capacity. If the individual FPR/TPR pairs are precisely known without sampling errors, they can be directly compared with the machine ROC curve; we interpret a human FPR/TPR pair below the machine ROC curve as evidence of the superiority of the machine algorithms. As shown in Figure 2, a human FPR/TPR pair below the machine ROC curve can be dominated by any point on

Figure 1. (Color online) Individual and Aggregate FPR/TPR Pairs



Notes. When doctors use a correctly specified propensity score model to process both public and private information variables, the collection of FPR/TPR pairs for human decision makers, represented by the triangular points, all lie above the machine ROC curve. However, the aggregate FPR/TPR of human decision makers can still lie below the machine ROC curve. See Section 2.5 for a more detailed discussion of how incentive heterogeneity can mask the nuances of comparing between machine learning and human decision making through an implication of Jensen's inequality.

Figure 2. (Color online) Population Human FPR/TPR Pair and the Machine ROC Curve



Notes. Point A matches the human FPR but has a higher TPR. Similarly, Point B matches the human TPR but has a lower FPR. The FPR and TPR of any point between points A and B on the machine ROC curve are smaller and larger, respectively, than the human decision maker's FPR/TPR pair.

the line segment of the ROC curve between A and B. Replacing the human FPR/TPR pair by any point on the A–B segment of the machine ROC curve segment results in higher TPR without increasing FPR or lower FPR without sacrificing TPR.

In reality, only an estimate of the FPR/TPR pair can be obtained from the empirical data. A statistical framework to compare the performance of algorithms with that of human decision makers requires accounting for estimation sampling errors. We focus on two main issues. First, we seek statistical evidence supporting the information superiority of an algorithm that favors the replacement of a human decision maker. Second, we aim to identify the most appropriate point on the machine ROC curve for this purpose. To address these two issues, we experiment with both a heuristic frequentist confidence set approach and a subsequent Bayesian inference approach. In both scenarios, confidence and credible regions are formed to inform the decision-making process. The Bayesian approach results in more replacement and additional improvement in the aggregate FPR/TPR pair. We suggest the Bayesian analysis as an applicable decision framework.

At a disaggregated level, each individual human can have either less or more information processing capacity than an algorithm has. For instance, Liang et al. (2019) propose a machine learning algorithm to diagnose childhood diseases that outperforms junior physician groups but marginally underperforms senior physician groups. Similar findings are reported in Peng et al. (2021). Esteva et al. (2017) and Kermany et al. (2018) also find substantial variability in the human experts' performance. Our framework chooses between each human decision maker and the algorithm for making future decisions. We identify a subset of human decision makers to be replaced by the algorithm, whereas the rest of human decision makers are retained.

We apply our statistical framework to a unique National Free Prepregnancy Checkups (NFPC) data set. The NFPC is a free health checkup service for conceiving couples across 31 provinces in China. In addition to the pregnancy outcome and numerous patient-case features, the data set also includes the doctors' IDs and diagnoses of adverse pregnancy outcomes. We first split the data into two parts. The first part is used to compare doctors with algorithms. Specifically, we employ a random forest (RF) method for diagnosing risky pregnancy, which achieves an area under the curve (AUC) above 0.68. Using a 95% credible level, our Bayesian approach suggests that the random forest algorithm outperforms 46.1% of doctors. The second part of the data set is used to evaluate the quality of the diagnosis procedure based on the combination of retained doctors and the algorithm. The combined decision-making procedure achieves an increase of 46.6% in the TPR and a reduction of 10.1% in the FPR. Additional detailed analysis further shows that the potential for machine algorithms to improve decision-making quality is substantial even when only a smaller portion of doctors are replaced. We also find that doctors practicing only in country hospitals have a lower replacement ratio compared with those having practiced in township clinics. The result of our Bayesian analysis suggests that 51.2% of doctors who have practiced only at township clinics are replaced compared with 39.8% of doctors who have practiced only at county hospitals.

1.1. Related Literature

Our paper relates to several strands of literature. An extensive literature studies the utilization of AI to enhance medical diagnostic capabilities. Machine learning, especially deep learning methods, which demonstrate great potential in tackling diverse and complex tasks, has been experimented to diagnose in multiple medical subfields, including ophthalmology (Kermany et al. 2018), cardiology (Rajpurkar et al. 2017, Hannun et al. 2019, Ren et al. 2021), dermatology (Esteva et al. 2017), respirology (Ardila et al. 2019, Liang et al. 2019), clinical psychology and psychiatry (Ashar et al. 2017, Yoon et al. 2022), and tumor and cancer detection (McKinney et al. 2020, Peng et al. 2021, Uhm et al. 2021, Tian et al. 2024). Additional surveys can be found in Erickson et al. (2017), Johnson et al. (2018), Dwyer et al. (2018), Chan et al. (2020), Huang et al. (2020), and Swanson et al. (2023).

Beyond medical diagnoses, AI is widely explored in social-economic decision making. See, for example, Berg et al. (2020), Sadhwani et al. (2021), and Fekadu et al. (2022) for loan approval and credit risk modeling; Fuster et al. (2019, 2022) and Vallee and Zeng (2019) for market structure implications of employing machine learning algorithms; Berk (2017), Zeng et al. (2017), and Kleinberg et al. (2018) for bail and parole analysis; and Chalfin et al. (2016) and Sajjadiani et al. (2019) for productivity assessment.

Numerous papers in the literature benchmark algorithms against human decision makers.[1] Whereas machine algorithms often outperform some aggregation of human decision makers, these findings are not uniform across individual human decision makers. Unlike machine algorithms, human decision makers can be highly heterogeneous in both incentives and information processing capacity. The health economics literature suggests that doctors may be swayed by multiple nonmedical incentives, including lawsuit avoidance (Studdert et al. 2005), financial gain (Johnson and Rehavi 2016), demand pressure from the patient (Lopez et al. 2018), and procedural skills (Currie and MacLeod 2017). Our analysis highlights the importance of modeling and restricting incentive heterogeneity in a human–AI comparison.

A growing literature also explores AI-assisted human decision making. Whereas Brennan et al. (2019), Han et al. (2020), Peng et al. (2021), Yang et al. (2021), and Tian et al. (2024) report salient performance improvement when doctors can refer to the prediction results of algorithms, Jin et al. (2024) and Yu et al. (2024) report contradicting evidence by which the performance of a portion of doctors is worsened by AI assistance. By offering doctors monetary incentives prior to the experiment, Wang et al. (2023, 2024) find that doctor diagnoses can be significantly influenced by both the machine information and the preset incentive structures. Whereas both Bansal et al. (2021) and Wang et al. (2023) find that giving doctors explanations increases the acceptance of machine recommendations, the effect of providing explanation on performance is inconclusive and varies. In the beverage vending machine business, a field experiment on product assortment decisions conducted by Kawaguchi (2021) finds heterogeneous adoption by workers dependent on location and both worker and machine characteristics. Iakovlev and Liang (2024) theoretically investigate the role of contextual and feature information between humans and AI.

Alur et al. (2023, 2024) propose both a test of whether human experts help improve machine algorithm predictions and a framework that integrates human expertise into algorithmic predictions. Our analysis also suggests that the human machine complementarity, argued for by Bansal et al. (2021) and Donahue et al. (2022), is not only theoretically challenging, but also difficult to implement empirically. The meta-analysis in Vaccaro et al. (2024) finds limited scope of human–AI effectiveness improvement. Overall, the mechanism by which AI assists human decision making is unclear. Our paper, thus, focuses on a simple strategy that identifies and replaces underperforming doctors. This modeling choice is also consistent with the analysis in Mullainathan and Obermeyer (2022) and Chan et al. (2022), who suggest that doctor performance cannot be explained by preference heterogeneity alone. In addition, the case-specific replacement algorithm described in Online Appendix A.6 does not outperform the baseline method for our NFPC data set.

A substantial body of literature discusses the discriminatory bias of algorithms (O'Neil 2017, Eubanks 2018, Berk et al. 2021). But the potential of AI and algorithmic decision making to reverse extant bias in the training data are also noted by Rambachan and Roth (2019), Rambachan et al. (2020), and Gillis et al. (2021). Through efficient information provision and scarce resource utilization, AI can be instrumental for inequality reduction. Using artificial intelligence to empower medical diagnoses in developing areas has been advocated by research including Adepoju et al. (2017), Guo and Li (2018), Trivedi et al. (2019), Mondal et al. (2023), and others. Bulathwela et al. (2021) discuss the comprehensive application of AI to narrow the education inequality gap. Our finding using the NFPC data set is also consistent with anecdotal information regarding the quality disparity across tiers of the hospital classification system in China.

Optimal decision making using machine learning techniques is also drawing increasing attention from econometrics. Kitagawa and Tetenov (2018), Mbakop and Tabord-Meehan (2021), Athey and Wager (2021), and Feng and Hong (2024) explore the asymptotic properties of empirical algorithm-based optimal decision making. Manski (2018) considers the identification problem when only the prediction conditional on a subset of covariates and the conditional distribution of the omitted covariates are known. In this paper, we assume, as in the baseline cases of Kitagawa and Tetenov (2018), that the machine algorithm is known and deterministic in view of the large data size and the comprehensive list of covariates in the NFPC data set.

The rest of this paper is organized as follows. Section 2 presents the statistical model of human–algorithm comparison. Section 3 describes the data and the algorithm. Section 4 reports results from the empirical analysis of the machine algorithm and human decision makers. Section 5 presents a set of synthetic data analysis for additional insights, and Section 6 concludes. Additional results are collected in the Online Appendix.

2. Replacing Human Decision Makers with Algorithms

Consider a sample data set with labels $Y_i \in \{0,1\}$ and features $X_i, i = 1, \ldots, n$. Define

$$\mathrm{TPR} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i \hat{Y}_i}{\hat{p}}, \quad \mathrm{FPR} = \frac{\frac{1}{n}\sum_{i=1}^{n} (1 - Y_i)\hat{Y}_i}{1 - \hat{p}},$$
$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{1}$$

where $\hat{Y}_i \in \{0,1\}$ is a predictor for Y_i based on a general decision rule such that $(X_i, Y_i, \hat{Y}_i)$ are independent and identically distributed (i.i.d.) draws from an underlying population. The law of large numbers implies that the FPR/TPR converge to their population analogs (denoted as β and α) as $n \to \infty$:

$$\beta = \frac{1}{p}\mathbb{E}Y_i\hat{Y}_i, \quad \alpha = \frac{1}{1-p}\mathbb{E}[(1 - Y_i)\hat{Y}_i],$$
$$\text{where} \quad p = \mathbb{E}Y_i = \mathbb{P}(Y_i = 1). \tag{2}$$

A key concept in this paper is incentive heterogeneity. In general terms, incentive refers to the motivation of certain actions, and incentive heterogeneity refers to different payoffs across multiple human decision makers for taking the same actions. Incentive heterogeneity can typically be classified intuitively as either intraindividual or interindividual. On the one hand, interindividual

incentive heterogeneity refers to situations in which the cost assessment only varies across decision makers but not across individual cases for each decision maker. On the other hand, intraindividual incentive heterogeneity refers to situations in which the cost assessments across individual cases for a given decision maker vary based on both publicly observed features x that are accessible by the machine algorithm and private information features u that are only available to human decision makers. To compare human decision makers and machines, we make the following two assumptions regarding incentive heterogeneity:

Assumption 1. *The machine ROC curve is generated by a propensity score model of prediction:*

$$\hat{y}_{i,M} = \mathbb{1}(m(x_i) > c).$$

The decision maker j's decision is also based on a perceived propensity score model:

$$\hat{y}_{i,j} = \mathbb{1}(q_j(x_{i,j}, u_{i,j}) > c_j).$$

Both $m(\cdot)$ and $q_j(\cdot)$ can be correctly specified or misspecified.

Assumption 1 rules out intraindividual incentive heterogeneity and allows us to focus on interindividual incentive heterogeneity. Doctors in our data set exhibit highly variable degrees of conservativeness in their diagnoses. The FPR of the doctors with at least 300 diagnoses has a mean of 0.236 and a standard deviation of 0.230. Some doctors tend to diagnose almost all patients as highly risky, and some doctors rarely make a positive diagnosis. The diffusive distribution of doctors' FPR/TPR pairs also indicates both information processing capacity difference and interindividual heterogeneity. The analysis in Chan et al. (2022) shows that, when only either diagnostic skill difference or interindividual incentive heterogeneity is present, the doctor FPR/TPR pairs will concentrate in a small region, which is not the case in our data. Importantly, Assumption 1 also allows for private information $u_{i,j}$ that is only observed by the doctors and not used in the machine learning algorithm. In the absence of the private information variable $u_{i,j}$, the doctors' diagnoses are perfectly predictable by the machine learning algorithm, which is very unlikely.

Assumption 2. *The machine propensity score model $m(x_i)$ and the jth decision maker's propensity score model $q_j(x_{i,j}, u_{i,j})$ are used without sampling errors in the machine decision making procedure and by the jth decision maker, respectively.*

The first requirement is justified by the fact that the machine learning model is implemented using a training data set that is orders of magnitude larger than the number of cases for each individual human decision maker. For instance, around 150,000 cases in the NFPC training data set are used to estimate the machine ROC curve, whereas the median number of diagnoses per doctor in the classification set is only approximately 400. The estimation error in the machine ROC curve is negligible relative to the error in estimating doctors' FPR/TPR pairs. The sampling error in estimating the machine ROC curve is studied in a subsequent paper by Feng and Hong (2024).

The second requirement that the $q_j(x_{i,j}, u_{i,j})$ are used without sampling errors by the jth decision maker is interpreted as stating that the behavior of the jth decision maker is consistent with the use of a propensity score model $q_j(x_{i,j}, u_{i,j})$. The method in this paper does not require the researcher to know the propensity score model $q_j(x_{i,j}, u_{i,j})$ for the jth decision maker and does not refer to any of the doctor propensity score models $q_j(x_{i,j}, u_{i,j}), j = 1, \ldots, J$. The second requirement also rules out decision makers learning from the sample cases. A doctor j does not update the doctor's propensity score model $q_j(x_{i,j}, u_{i,j})$ during the data sample period.

This is a reasonable assumption in our empirical data set because the eventual pregnancy outcomes are typically not available to doctors when they diagnose consecutive patient cases. We are not concerned with how doctors procure knowledge to develop their diagnosis model $q_j(x_{i,j}, u_{i,j})$. Doctor j might procure $q_j(x_{i,j}, u_{i,j})$ from trainings obtained from a medical school or from working with a large pool of patients prior to the sampling period. Section 3 provides more details about the data-processing procedure, in which sampling errors arise mainly from the relatively fewer number of observations for each individual decision maker.

2.1. Human FPR/TPR Pairs and ROC Curves in the Population

The parameter of interest is the population pair of true positive and false positive rates for an individual decision maker: $\theta_H = (\alpha_H, \beta_H)$, where α and β are defined as in (2) and the subscript H refers to "human." If θ_H is above a machine ROC curve, as in Figure 1, then given the FPR α_H, the algorithm has a lower TPR β_M than humans have. Similarly, given the TPR β_H, the machine has a larger FPR α_M than humans have. In this sense, humans outperform the algorithm. In particular, under Assumption 1, if humans correctly use at least as much information as the algorithm does, in the sense that the propensity model is correctly specified, then θ_H lies above the machine ROC. The formal argument is given in Online Lemma A.1.

Figure 2 shows a θ_H pair for a human decision maker that is below the machine ROC curve. A machine decision rule corresponding to a point on the machine ROC curve between A and B performs better than the human decision maker. The segment of the curve to the left of B (to the right of A) has a smaller α_M and a smaller β_M (a larger β_M but a larger α_M) than the human decision

maker. These segments of the machine ROC curve are not directly comparable to the human decision maker.

The classic Neyman–Pearson lemma states that, when $p(x) = \mathbb{P}(Y = 1|x)$ is correctly specified, for any pair of positive numbers (η, ϕ), there exists a cutoff value c such that the point (α_c, β_c) on the ROC curve corresponding to the decision rule $\hat{y} = \mathbb{1}(p(x) > c)$ maximizes a linear combination $\phi\beta - \eta\alpha$ of the TPR and FPR. Conversely, each point on this ROC curve optimizes some linear combination of TPR and FPR. The Neyman–Pearson lemma is also motivated by a Bayesian decision maker who minimizes posterior expected loss of misclassification upon observing the features x based on a loss matrix of the following form:

Loss matrix	$\hat{y} = 0$	$\hat{y} = 1$
$y = 0$	0	c_{01}
$y = 1$	c_{10}	0,

where c_{10} represents the cost of misclassifying one as zero and c_{01} represents the cost of misclassifying zero as one. Each point on the ROC curve, corresponding to a decision rule $\hat{y} = \mathbb{1}(p(x) > c)$, minimizes Bayesian posterior expected loss for some choice of c_{10} and c_{01} through the relation $c = c_{01}/(c_{10} + c_{01})$. Importantly, the cost matrix c_{01}, c_{10} and linear combination coefficients η, ϕ in the Neyman–Pearson lemma are independent of the features x in order to provide optimality justification of the points on the ROC curve. Assumption 1 is then interpreted as each doctor behaving as a Bayesian decision maker adopting the same constant cost matrix c_{10} and c_{01} among the doctor's patients. The ex ante expected minimized cost for the Bayesian decision maker corresponds to a linear combination of FPR and TPR:

$$\begin{aligned} &c_{10}\mathbb{E}[p(X)(1-\hat{Y})] + c_{01}\mathbb{E}[(1-p(X))\hat{Y}] \\ &= (1-p)c_{01}\mathrm{FPR} - pc_{10}\mathrm{TPR} + pc_{10}. \end{aligned}$$

If we denote points A and B in Figure 2 as (α_H, β_R) and (α_R, β_H), then any point (α_M, β_M) between (α_R, β_H) and (α_H, β_R) on the machine ROC curve satisfies $\alpha_M < \alpha_H$ and $\beta_M > \beta_H$. For arbitrary $\phi, \eta > 0$ and c_{10}, c_{01},

$$\begin{aligned} &\eta\alpha_M - \phi\beta_M < \eta\alpha_H - \phi\beta_H, \quad \text{and} \\ &(1-p)c_{01}\alpha_M - pc_{10}\beta_M < (1-p)c_{01}\alpha_H - pc_{10}\beta_H. \end{aligned}$$

Therefore, compared with human decisions, any point (α_M, β_M) between (α_R, β_H) and (α_H, β_R) on the ROC curve achieves a better linear combination for all possible values of ϕ, η, and lower expected cost for all loss matrixes of c_{01}, c_{10}. In this sense, (α_M, β_M) dominates (α_H, β_H).

Recall that θ_H is a population parameter that measures the expected performance of the human decision maker. When θ_H lies below the machine ROC curve, on average, the machine algorithm can outperform the human decision maker by increasing the population TPR or reducing the population FPR. However, unless the algorithm can make predictions with perfect accuracy, for each specific patient case, the human decision maker can either underperform or outperform the algorithm.

Our framework is likely to be applicable when people make many repeated decisions to allow for a sufficiently precise estimate of their error rates. It is less applicable in situations when people only ever make a few decisions of a type, a setting that may arise even when decisions are frequent if some action rarely is or should be taken. The lack of information renders the task of distinguishing humans from algorithms difficult even if humans might be detectably worse in aggregate.[2] Our application features a comprehensive set of covariates, a large data set, reliable and measurable outcomes, and a stable prediction environment that are consistent with the scenarios in the important papers by Kahneman and Klein (2009) and Currie and MacLeod (2017).

2.2. Human FPR/TPR Pair and ROC Curve Comparison in the Sample

Typically, the population value of θ_H is not directly observed and, instead, needs to be estimated from a data set. Denote the sample estimates (FPR, TPR) in (1) as $\hat{\theta}_H = (\hat{\alpha}_H, \hat{\beta}_H)$. In the presence of sampling uncertainty, the inference problem pertains to how we can make a probabilistic statement regarding whether θ_H is above or below the machine ROC curve. From a frequentist analysis point of view, because the sample estimate $\hat{\theta}_H$ is a vector function of a multinomial distribution, the finite sample distribution function of $\hat{\theta}_H$ can be analytically intractable. Large sample frequentist analysis and simulation methods, however, are facilitated by the joint asymptotic normal distribution of $\hat{\theta}_H$.

Lemma 1. *For an i.i.d. sample $\{Y_i, \hat{Y}_i\}_i^n$, the joint asymptotic distribution of a human FPR/TPR pair $\hat{\theta}_H = (\hat{\alpha}_H, \hat{\beta}_H)$ is multivariate normal. In particular,*

$$\sqrt{n}(\hat{\theta}_H - \theta_H) \xrightarrow{d} N(0, \Sigma), \quad \Sigma = \begin{pmatrix} \frac{\alpha_H(1-\alpha_H)}{1-p} & 0 \\ 0 & \frac{\beta_H(1-\beta_H)}{p} \end{pmatrix}.$$

The proof of Lemma 1 is given in the Online Appendix and follows from standard arguments for approximating multinomial distributions with normal limits. The asymptotic covariance Σ can be consistently estimated by a sample analog $\hat{\Sigma}$, where the parameters α_H, β_H, and p are replaced by $\hat{\alpha}_H, \hat{\beta}_H$, and $\hat{p}$. The resulting asymptotic confidence set of the FPR/TPR pairs is elliptically shaped. As illustrated in Figure 3(a), the

ellipse indicates a 95% level confidence set for a decision maker in our data set.

An immediate consequence of Lemma 1 is that a direct comparison of the sample estimate $(\hat{\alpha}_H, \hat{\beta}_H)$ with the machine ROC curve is likely to be insufficient. Consider a situation in which the true population pair of (α_H, β_H) lies strictly below the machine ROC curve: $\beta_H < g(\alpha_H)$, where $g(\cdot)$ denotes the machine ROC curve. Then, with probability converging to one, $(\hat{\alpha}_H, \hat{\beta}_H)$ lies strictly below the machine ROC. Suppose we replace $(\hat{\alpha}_H, \hat{\beta}_H)$ by $(\hat{\alpha}_H, g(\hat{\alpha}_H))$. By Lemma 1, $\mathbb{P}(\hat{\alpha}_H \geqslant \alpha_H) \to 0.5$. In other words, the replacement point $(\hat{\alpha}_H, g(\hat{\alpha}_H))$ on the ROC curve does not dominate the true population pair of (α_H, β_H) approximately half of the time. A similar situation arises if we replace $(\hat{\alpha}_H, \hat{\beta}_H)$ by $(g^{-1}(\hat{\beta}_H), \hat{\beta}_H)$. In order to provide a probabilistic statement regarding the property of the replacement procedure, a more formal statistical framework is necessary. We consider both a heuristic frequentist approach and a Bayesian approach.

2.3. A Heuristic Frequentist Approach

We initially experiment with a heuristic procedure based on the frequentist principle, the predominant school of thought in statistical inference which typically begins with confidence set construction. The procedure is described as follows:

- Form a confidence set $\hat{S}$ for θ_H. By Lemma 1, $n(\hat{\theta}_H - \theta_H)^T \hat{\Sigma}^{-1} (\hat{\theta}_H - \theta_H)$ converges to a chi-square distribution with two degrees of freedom, where $\hat{\Sigma}$ is any consistent estimator of Σ. A confidence set on this asymptotic distribution is elliptically shaped.
- For each $s = (\alpha_s, \beta_s) \in \hat{S}$, define A_s as the set of points that dominate s:

$$A_s = \{\theta = (\alpha, \beta) : \alpha \leqslant \alpha_s, \beta \geqslant \beta_s\}. \tag{3}$$

- Define $A = \cap_{s \in \hat{S}} A_s$: A is the set of points that dominate all the points in $\hat{S}$.
- Next, define $\overline{A} = A \cap \text{ROC}$: $\overline{A}$ is the set of points on the machine ROC curve that simultaneously dominate all the points in $\hat{S}$.

The convention of choosing a large confidence level suggests that the status quo of human decision making is to be replaced by an algorithm only when there is overwhelming evidence from the data indicating the superiority of the algorithm. In our application, we only consider the machine algorithm to outperform the human decision when $\overline{A} \neq \emptyset$, that is, when there exists a point on the machine ROC curve that dominates an entire confidence set of the human decision maker's θ_H parameter. In essence, we use estimates of the sampling uncertainty of the FPR/TPR pair of $\hat{\theta}_H$ to provide guidance on how far $\hat{\theta}_H$ needs to be below the machine ROC curve in order to justify replacing the human decision maker with the machine algorithm.

In the elliptical confidence set depicted in Figure 3(a), denote by $(\alpha_{\beta_{high}}, \beta_{high})$ the point that achieves the highest TPR on the confidence set, and denote by $(\alpha_{low}, \beta_{\alpha_{low}})$ the point that achieves the smallest FPR on the confidence set. The point P in Figure 3(a) corresponds to $(\alpha_{low}, \beta_{high})$. There are three possibilities for the relation between the position of the confidence set for the human FPR/TPR pair θ_H and the machine ROC curve.

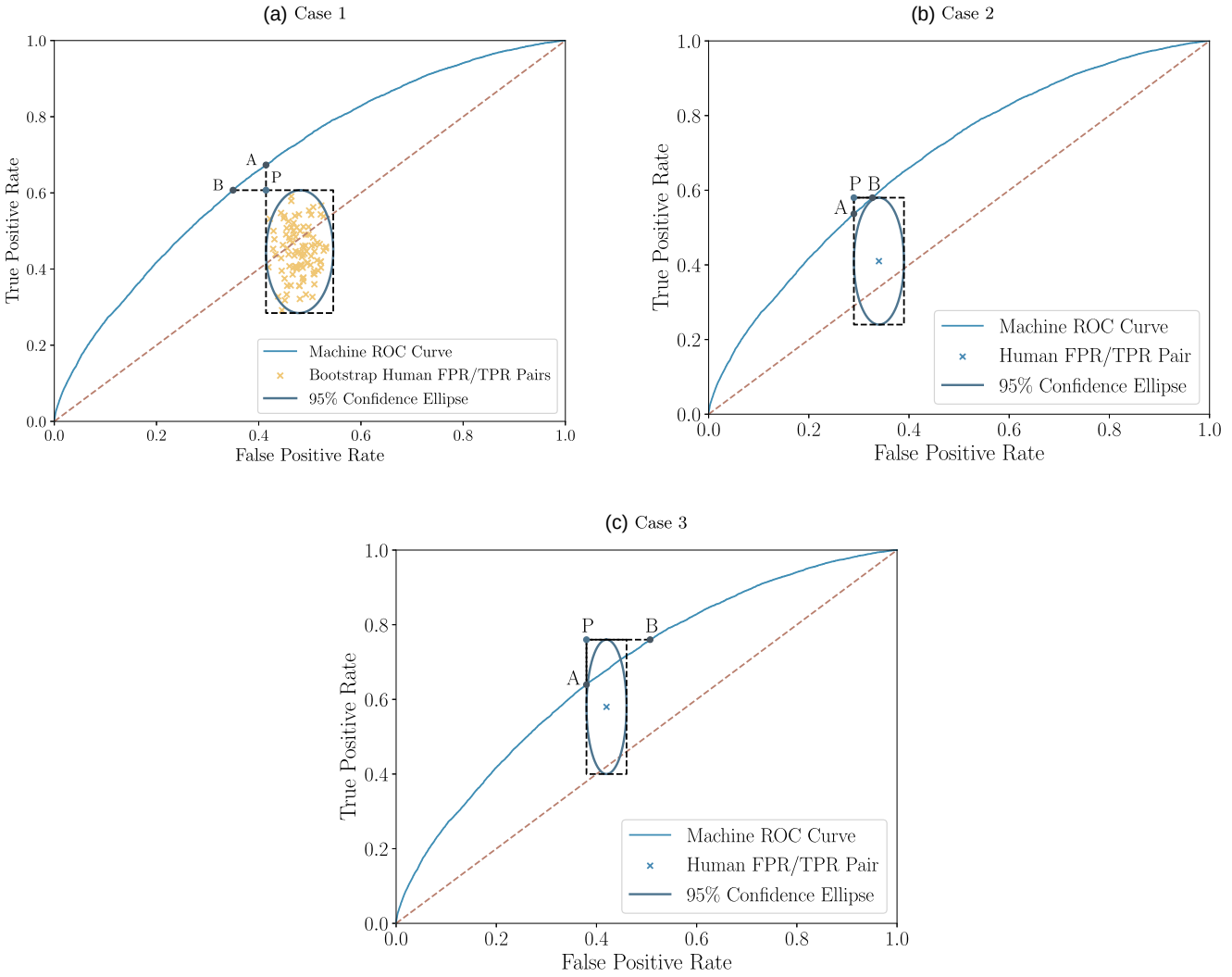
1. Case 1 (Figure 3(a)): The human confidence set and point $P = (\alpha_{low}, \beta_{high})$ are both below the machine ROC curve. We consider the human decision maker to perform worse than the algorithm and, hence, can be replaced by the algorithm. On the machine ROC curve, we can find two points, namely, (α_R, β_{high}) and (α_{low}, β_R), labeled as points B and A in Figure 3(a), such that any point on the ROC curve between A and B can provide a machine decision rule to replace human decision making.
2. Case 2 (Figure 3(b)): The entire human confidence set is below the ROC curve, but the point $P = (\alpha_{low}, \beta_{high})$ is above the ROC curve. In this case, even though each point in the confidence region of θ_H is dominated by some point on the machine ROC curve, it is not possible to find a nonempty fraction of the machine ROC curve that simultaneously dominates all the points in the human confidence set. The data evidence is not sufficiently convincing to account for the randomness of estimating θ_H by $\hat{\theta}_H$. Consequently, the human decision maker is not replaced by the algorithm.
3. Case 3 (Figure 3(c)): The human confidence set has a certain area above the machine ROC curve. The human decision maker is not replaced by the algorithm either.

In summary, if the point $P = (\alpha_{low}, \beta_{high})$ is below the ROC curve, the human decision maker is replaced; otherwise, the human decision maker is not. In cases 2 and 3, it is possible that the human decision maker is sufficiently capable compared with the machine algorithm. It is also possible that the human θ_H is not estimated precisely enough, for example, because of a lack of historical data that results in a large confidence set.

In frequentist statistical inference, θ_H is a fixed number. It is either above or below the machine ROC curve. The confidence set itself $\hat{S}$ is a random set such that $\mathbb{P}(\theta_H \in \hat{S}) \to 1 - \alpha$. For a conventional size of $\alpha = 0.05$, if the confidence set is to be constructed 100 times from different samples, $\hat{S}$ contains θ_H approximately 95 times. Consequently, by construction, $\overline{A}$ is a random set such that asymptotically,

$$\begin{aligned}\mathbb{P}(\theta_H \in \hat{S}) = {} & \mathbb{P}\{\overline{A} = \emptyset \text{ and P dominates } \theta_H\} \\ & + \mathbb{P}\{\overline{A} \neq \emptyset \text{ and } \overline{A} \text{ dominates } \theta_H\} \geqslant 1 - \alpha.\end{aligned}$$

Figure 3. (Color online) Three Cases of the Heuristic Approach



Notes. The heuristic frequentist approach makes use of the point $\mathrm{P} := (\alpha_{low}, \beta_{high})$ that corresponds to the smallest FPR and the highest TPR of the elliptical confidence set. Panel (a) shows the case in which *P* lies below the machine ROC curve and there exist dominating points on the curve. The crosses are the bootstrapped FPR/TPR pairs. Panel (b) shows the case in which the confidence set lies below the machine ROC curve, but there is no dominating point on the curve. Panel (c) shows the case in which a portion of the confidence set lies above the machine ROC curve.

Therefore, we have, asymptotically,

$$\mathbb{P}\{\overline{A} \text{ dominates } \theta_H | \overline{A} \neq \emptyset\} \geqslant \frac{1-\alpha-\mathbb{P}\{\overline{A}=\emptyset \text{ and P dominates } \theta_H\}}{\mathbb{P}\{\overline{A}\neq\emptyset\}}.$$

If θ_H lies below the ROC curve, $\mathbb{P}\{\overline{A}=\emptyset\} \to 0$ as the number of cases for the human decision maker increases to infinity. Furthermore, $\mathbb{P}\{\overline{A}=\emptyset \text{ and P dominates } \theta_H\} < \mathbb{P}\{\overline{A}=\emptyset\} \to 0$, implying that, asymptotically, $\mathbb{P}\{\overline{A} \text{ dominates } \theta_H | \overline{A}\neq\emptyset\} \geqslant 1-\alpha$. If θ_H lies above the ROC, we can control $\mathbb{P}\{\overline{A}=\emptyset\} \geqslant 1-\alpha$ by reasoning that

$$\mathbb{P}\{\overline{A}=\emptyset\} = \mathbb{P}\{\mathrm{P} \text{ is above ROC}\} \geqslant \mathbb{P}\{\mathrm{P} \text{ dominates } \theta_H\} \geqslant \mathbb{P}(\theta_H \in \hat{S}) \geqslant 1-\alpha.$$

However, the finite sample size of the test can be severely distorted. By insisting on finding a segment of the machine ROC curve that simultaneously dominates the entire confidence region of θ_H, the frequentist procedure is conservative by construction.

Whereas conventional confidence sets are mostly symmetric, it is well known in statistics that, by inverting a hypothesis test, alternative confidence sets can be constructed as the collection of parameter values for which the null hypothesis is not rejected. For example, if we let the null hypothesis be $H_0 : \theta_H = \theta_{0H}$ and the alternative hypothesis be $H_1 : \alpha_H > \alpha_{0H}$ or $\beta_H < \beta_{0H}$, a test can be formed to reject the null hypothesis H_0 when either $\hat{\alpha}_H > \alpha_H + c$ or when $\hat{\beta}_H < \beta_H - d$, where c and d are chosen by the size of the test. The confidence set of θ_{0H} in which the test does not reject takes the form of $\{\theta_{0H} : \alpha_{0H} \geqslant \hat{\alpha}_H - c, \text{ and } \beta_{0H} \leqslant \hat{\beta}_H + d\}$, which is lower rectangular. Whereas forming confidence sets based on inverting inequality tests is theoretically intriguing, it is not often used empirically.

An alternative in Online Appendix A.3 is to test whether the population value of θ_H is above or below the machine ROC curve. We do not use this test because of its limitation that, even if the null hypothesis of the status quo of human decision making is rejected, it does not locate a point of the machine ROC curve to replace human decisions as discussed in Section 2.2.

2.4. A Bayesian Approach

In the following, we focus on an alternative Bayesian method, which often allows for a more transparent interpretation than frequentist inference. The Bayesian approach combines a prior distribution of the underlying population parameters with the likelihood of the data given the population parameters to produce a posterior distribution of the parameters given the data, which is then used in a decision-theoretic framework in which a threshold is chosen to minimize the posterior expected loss function. The combination of a multinomial likelihood function with a Dirichlet conjugate prior drastically improves computability. Simulating the posterior distribution of the key FPR/TPR parameters to construct a Bayesian credible set is facilitated by writing them as a vector function of the multinomial parameters whose posterior distribution is given analytically through the conjugate prior. The empirical results suggest that the Bayesian method is less conservative and suggests replacement of more doctors by the machine algorithm than the frequentist method does.

Computing a Bayesian posterior distribution requires two key inputs: the likelihood of the data given the population value of θ_H, and a prior distribution for the underlying population parameters. Each observation in the data set consists of both the doctor diagnoses and the ground truth of pregnancy outcomes and follows a multinomial distribution with four categories: diagnosed as risky and abnormal birth outcome, not diagnosed as risky and abnormal birth outcome, diagnosed as risky and normal birth outcome, not diagnosed as risky and normal birth outcome. The parameters of the multinomial probabilities are

$$t_1 = \mathbb{E}Y\hat{Y},\ t_2 = \mathbb{E}[(1-Y)\hat{Y}],\ t_3 = \mathbb{E}[Y(1-\hat{Y})],$$
$$t_4 = \mathbb{E}[(1-Y)(1-\hat{Y})].$$

Because $t_1 + t_2 + t_3 + t_4 = 1$, we choose the first three as free parameters. The multinomial distribution is a completely specified parametric model and allows for exact likelihood Bayesian posterior distribution computation. The likelihood of the data given the free parameters $\boldsymbol{t} = (t_1, t_2, t_3)$ depends only on a set of sufficient statistics $\hat{\boldsymbol{t}} = (\hat{t}_1, \hat{t}_2, \hat{t}_3)$, where

$$\hat{t}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i\hat{Y}_i, \quad \hat{t}_2 = \frac{1}{n}\sum_{i=1}^{n}[(1-Y_i)\hat{Y}_i],$$
$$\hat{t}_3 = \frac{1}{n}\sum_{i=1}^{n}[Y_i(1-\hat{Y}_i)]. \tag{4}$$

In particular, we can write the data likelihood as (in which $\mathbb{D}$ refers to data):

$$L(\mathbb{D}|t) = \binom{n}{n\hat{t}_1} t_1^{n\hat{t}_1} \binom{n-n\hat{t}_1}{n\hat{t}_2} t_2^{n\hat{t}_2} \binom{n-n\hat{t}_1-n\hat{t}_2}{n\hat{t}_3} t_3^{n\hat{t}_3} t_4^{n(1-\hat{t}_1-\hat{t}_2-\hat{t}_3)}. \tag{5}$$

Given a prior distribution of t, denoted as $\pi(t)$, the posterior distribution of the parameter t, denoted as $p(t|\mathbb{D})$ can usually be simulated or calculated analytically:

$$p(t|\mathbb{D}) = \frac{\pi(t)L(\mathbb{D}|t)}{\int \pi(t')L(\mathbb{D}|t')\mathrm{d}t'}.$$

We are interested in the posterior distribution of $\theta_H = (\alpha_H, \beta_H)$, which can be written as functions of the underlying multinomial parameters: $h(t) = (\alpha_H(t), \beta_H(t))$. If we can simulate from the posterior distribution $p(t|\mathbb{D})$ and denote the simulated draws as $t_r, r = 1, \ldots, R$, then the posterior distribution of θ_H, denoted as $p(\theta_H|\mathbb{D})$, can be estimated using the empirical distribution of $h(t_r), r = 1, \ldots, R$.

Computing the posterior distribution using a multinomial likelihood is facilitated by the Dirichlet conjugate prior. A Dirichlet prior on the K-dimensional simplex of $t_k, k = 1, \ldots, K$, where $\sum_{k=1}^{K} t_k = 1$, is specified by hyperparameters $\boldsymbol{\gamma} = (\gamma_k, k = 1, \ldots, K)$. Its density is given by $\pi(t|\boldsymbol{\gamma}) = \prod_{k=1}^{K} t_k^{\gamma_k - 1}/B(\boldsymbol{\gamma})$, where $B(\gamma)$ is the multivariate Beta function. See, for example, Kotz et al. (2019). The case with $\gamma_k = \gamma$, $\forall k = 1, \ldots, K$ is called a symmetric Dirichlet distribution. In one dimension, in which $K = 1$, the Dirichlet distribution specializes to a Beta distribution, which, in turn, includes the uniform distribution as a special case when $\gamma = 1$.

The posterior distribution is also Dirichlet with parameters $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_k, k = 1, \ldots, K)$, where $\hat{\gamma}_k = \gamma_k + n\hat{t}_k$ for each $k = 1, \ldots, K$. Simulating from the posterior distribution of $\theta_H = (\alpha_H, \beta_H)$ can be accomplished by recomputing $\theta_{H,r} = h(\boldsymbol{t}_r)$ after drawing for t_r from the posterior Dirichlet distribution with parameters $\hat{\boldsymbol{\gamma}}$. Specifically, for each $\boldsymbol{t}_r = (t_{r,k})_{k=1}^{K}$, we calculate

$$\theta_{H,r} = h(\boldsymbol{t}_r) = \Big(\alpha_H(\boldsymbol{t}_r),\ \beta_H(\boldsymbol{t}_r)\Big), \quad \text{where}$$
$$\alpha_H(\boldsymbol{t}_r) = \frac{\boldsymbol{t}_{r,2}}{\boldsymbol{t}_{r,2} + \boldsymbol{t}_{r,4}},\ \beta_H(\boldsymbol{t}_r) = \frac{\boldsymbol{t}_{r,1}}{\boldsymbol{t}_{r,1} + \boldsymbol{t}_{r,3}}. \tag{6}$$

The simulated draws $\theta_{H,r}, r = 1, \ldots, R$ can be used to estimate posterior probabilities. For example, the posterior probability that θ_H lies below the ROC curve, denoted as

$$\int_{\theta_H \text{ below ROC}} p(\theta_H|\mathbb{D})\mathrm{d}\theta_H, \tag{7}$$

can be estimated by the fraction of times in which $\theta_{H,r}$ lies below the ROC curve:

$$\frac{1}{R}\sum_{r=1}^{R} \mathbb{1}(\theta_{H,r} \text{ lies below the machine ROC curve})$$
$$= \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}(\beta_H(\boldsymbol{t}_r) \leqslant g(\alpha_H(\boldsymbol{t}_r))).$$

We are mainly interested in the maximum posterior probability of θ_H that are dominated simultaneously by a single point on the ROC curve, denoted as

$$q_{max} \equiv \max_{\alpha\in[0,1]} \int_{\alpha_H \geqslant \alpha, \beta_H \leqslant g(\alpha)} p(\theta_H|\mathbb{D})\mathrm{d}\theta_H, \tag{8}$$

and the corresponding maximizing point, denoted as $\theta_{max} = (\alpha_{max}, \beta_{max} = g(\alpha_{max}))$, where

$$\alpha_{max} \equiv \arg\max_{\alpha\in[0,1]} \left\{ \int_{\alpha_H \geqslant \alpha, \beta_H \leqslant g(\alpha)} p(\theta_H|\mathbb{D})\mathrm{d}\theta_H \right\}. \tag{9}$$

Both (8) and (9) can be estimated by the simulated posterior draws of $\theta_{H,r}, r = 1, \ldots, R$ by

$$\hat{q}_{max} = \max_{\alpha\in[0,1]} \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}(\alpha_H(\boldsymbol{t}_r) \geqslant \alpha, \beta_H(\boldsymbol{t}_r) \leqslant g(\alpha)),$$
$$\hat{\alpha}_{max} = \arg\max_{\alpha\in[0,1]} \sum_{r=1}^{R} \mathbb{1}(\alpha_H(\boldsymbol{t}_r) \geqslant \alpha, \beta_H(\boldsymbol{t}_r) \leqslant g(\alpha)). \tag{10}$$

More generally, decision-theoretic Bayesian analysis minimizes posterior expected losses. Let $\rho(\theta_M, \theta_H)$ denote a loss function that represents the disutility of replacing a human FPR/TPR pair θ_H by a machine FPR/TPR pair θ_M. Consistent with Assumptions 1 and 2, we specify $\rho(\theta_M, \theta_H) = 0$ when $\alpha_M \leqslant \alpha_H$ and $\beta_M \geqslant \beta_H$ or when the machine FPR/TPR pair strictly dominates the human FPR/TPR pair. Otherwise, $\rho(\theta_M, \theta_H)$ may be strictly positive. Given the choice of $\rho(\theta_M, \theta_H)$, a point on the ROC curve can be chosen to minimize the posterior expected loss

$$\theta_M^0 = \arg\min_{\theta_M:\beta_M = g(\alpha_M)} \int \rho(\theta_M, \theta_H) p(\theta_H|\mathbb{D})\mathrm{d}\theta_H, \tag{11}$$

which is also estimated using the simulated posterior draws of $\theta_{H,r}, r = 1, \ldots, R$ as

$$\hat{\theta}_M^0 = \arg\min_{\theta_M:\beta_M = g(\alpha_M)} \frac{1}{R}\sum_{r=1}^{R} \rho(\theta_M, \theta_{H,r}).$$

In particular, (8) is a special case of (11) when $\rho(\theta_M, \theta_H) = 1 - \mathbb{1}(\alpha_H \geqslant \alpha_M, \beta_H \leqslant \beta_M)$. It is also possible to explore more general loss functions. For example, we may weight the loss of replacing θ_H by θ_M by the distance between θ_H and the machine ROC curve:

$$\rho(\theta_M, \theta_H) = 1 - \mathbb{1}(\alpha_H \geqslant \alpha_M, \beta_H \leqslant \beta_M) \cdot \inf_{\theta}\{\|\theta_H - \theta\| : \theta \text{ on ROC}\}, \tag{12}$$

where $\|\cdot\|$ is the Euclidean norm. Intuitively, the farther a human FPR/TPR pair θ_H is from the ROC curve, the smaller the loss (or the larger the benefit) of replacing it with a dominating point on the machine ROC curve. Alternatively, we can replace the distance from θ_H to the ROC curve by the distance of each θ_M to the complement of the set of FPR/TPR points that dominate θ_H and use the loss function:

$$\rho(\theta_M, \theta_H) = 1 - \mathbb{1}(\alpha_H \geqslant \alpha_M, \beta_H \leqslant \beta_M) \cdot \min(\alpha_H - \alpha_M, \beta_M - \beta_H). \tag{13}$$

The Euclidean distance from the machine ROC curve may also not fully capture the loss of not replacing a human FPR/TPR pair θ_H that lies under the machine ROC curve. For example, a diagonal ROC curve corresponds to completely randomized decision making without using feature information. Similarly, a ROC curve below the diagonal can be generated by a decision rule that awards rather than penalizes misclassification errors. These considerations suggest that replacing a human FPR/TPR pair θ_H below the diagonal line with a dominating θ_M along the machine ROC curve should not entail losses. We can adopt a convention of normalizing the loss to be one when θ_M does not dominate θ_H, that is, when $\alpha_M \geqslant \alpha_H$ or $\beta_M \leqslant \beta_H$.

We then specify the loss $\rho(\theta_M, \theta_H)$ when θ_M dominates θ_H and when θ_H lies between the machine ROC curve and the diagonal line. Each of such θ_H can be reproduced by a convex combination of a point on the machine ROC curve and a point on the diagonal line, using a decision rule that randomizes between the point on the machine ROC curve and the point on the diagonal. For example, $(\alpha_H, g(\alpha_H))$ is generated by a machine-based decision rule $\hat{Y}_i = \mathbb{1}(m(X_i) > c_0)$ for some threshold c_0, while (α_H, α_H) is generated by a fully randomized decision $\hat{Y}_i = \mathbb{1}(U_i \geqslant 1 - \alpha_H)$, where U_i is independently uniformly distributed on $(0, 1)$. Then, (α_H, β_H) can be generated by a lottery that places weight $(\beta_H - \alpha_H)/(g(\alpha_H) - \alpha_H)$ on $\hat{Y}_i = \mathbb{1}(m(X_i) > c_0)$ and the remaining weight on $\hat{Y}_i = \mathbb{1}(U_i \geqslant 1 - \alpha_H)$. More precisely, if V_i denotes an independent random variable uniformly distributed on $(0, 1)$, the decision rule

$$\hat{Y}_i = \mathbb{1}\left(V_i \leqslant \frac{\beta_H - \alpha_H}{g(\alpha_H) - \alpha_H}\right) \mathbb{1}(m(X_i) > c_0) + \mathbb{1}\left(V_i > \frac{\beta_H - \alpha_H}{g(\alpha_H) - \alpha_H}\right) \mathbb{1}(U_i > 1 - \alpha_H)$$

generates the (α_H, β_H) FPR/TPR pair. The larger the

weight $\lambda_{\theta_H} \equiv (\beta_H - \alpha_H)/(g(\alpha_H) - \alpha_H)$ placed on the machine ROC curve, the smaller the benefit or the larger the loss of replacing θ_H with a dominating pair θ_M on the machine ROC curve. The randomization scheme motivates using the weight λ_{θ_H} as a component of the loss function

$$\rho(\theta_M, \theta_H) = 1 - \mathbb{1}(\alpha_H \geqslant \alpha_M, \beta_H \leqslant \beta_M)(1 - \max\{\lambda_{\theta_H}, 0\}). \tag{14}$$

Other randomization schemes can also be used. We refer to the above construction as a vertical decomposition. Similarly, a horizontal decomposition replaces the weights by $\lambda_{\theta_H} \equiv (\beta_H - \alpha_H)/(\beta_H - g^{-1}(\beta_H))$. In addition, we experiment with decomposing the distance between $\theta_M = (\alpha_M, \beta_M)$ and the diagonal line into two parts. The first part is the distance from θ_M to the boundary of the set of FPR/TPR points that do not dominate θ_H. The second part is the distance from this boundary to the diagonal line. The corresponding loss function that mirrors is

$$\rho(\theta_M, \theta_H) = 1 - \mathbb{1}(\alpha_H \geqslant \alpha_M, \beta_H \leqslant \beta_M)(1 - \max\{\lambda_{\theta_H, \theta_M}, 0\}), \tag{15}$$

where distance can be measured either horizontally, implying that $\lambda_{\theta_H, \theta_M} = (\beta_M - \alpha_H)/(\beta_M - \alpha_M)$, or vertically, implying that $\lambda_{\theta_H, \theta_M} = (\beta_H - \alpha_M)/(\beta_M - \alpha_M)$.

When $\alpha_M < \alpha_H$ and $\beta_M > \beta_H$, replacing θ_H by θ_M is beneficial. The benefit, $b(\theta_M, \theta_H)$, is likely to be larger when θ_H is farther below the machine ROC curve. When either $\alpha_M > \alpha_H$ or $\beta_M < \beta_H$, we expect a cost $\ell(\theta_M, \theta_H)$. A general loss function consists of both a cost component and a benefit component:

$$\begin{aligned}\rho(\theta_M, \theta_H) = {} & \mathbb{1}(\alpha_H < \alpha_M \text{ or } \beta_H > \beta_M)\ \ell\ (\theta_M, \theta_H) \\ & - \mathbb{1}(\alpha_H \geqslant \alpha_M, \beta_H \leqslant \beta_M)\ b\ (\theta_M, \beta_H).\end{aligned} \tag{16}$$

For each loss function, a guardrail threshold is used to account for the status quo of human decision making. A human decision maker is only replaced by the machine algorithm when the posterior expected loss of the replacement is below a given prespecified level or when the posterior expected benefit exceeds a certain level.

In addition to the deterministic decision rules of relying exclusively on either an individual doctor or the machine algorithm, randomized decision rules can be generated by mixing between a given point on the machine ROC curve and an individual doctor. Let ω denote the randomizing probability of using the machine algorithm and $1 - \omega$ the corresponding probability of relying on the human decision makers. The minimized expected posterior loss of the randomized decision rule can be written as

$$\min_{\omega \in [0,1],\, \theta_M : \beta_M = g(\alpha_M)} \int \rho(\omega\theta_M + (1-\omega)\theta_H, \theta_H)\, p\,(\theta_H | \mathbb{D}) \mathrm{d}\theta_H. \tag{17}$$

Without expert domain knowledge of medical diagnoses, it is difficult to implement (16) and (17). Investigating expert domain knowledge is beyond the scope of the current paper.

Our decision framework differs conceptually from the minimization of posterior expected loss function in Bayesian point estimation. A Bayesian point estimator of θ_H, denoted as $\hat{\theta}_{H,B}$, is often defined as the minimizer of an expected posterior loss function:

$$\hat{\theta}_{H,B} = \arg \min_{\theta \in [0,1]^2} \int \ell(\theta - \theta_H) p(\theta_H | \mathbb{D}) \mathrm{d}\theta_H.$$

The estimation loss function $\ell(\theta - \theta_H)$ typically depends only on $|\theta - \theta_H|$ and is increasing in $|\theta - \theta_H|$, which is different from our decision loss function $\rho(\theta_M, \theta_H)$.

2.5. Discussion: Incentives and Costs

The visual illusion shown in Figure 1 that hampers the comparison between machine learning algorithms and human decision makers is an immediate artifact of Jensen's inequality. An optimal ROC generated by a correctly specified propensity score function is necessarily concave because it is the primal function of a concave program. See, for example, Feng et al. (2022). For the jth doctor, denote the FPR as $\alpha_j = \mathbb{E}(1 - Y_{ij})\hat{Y}_{ij}/(1 - p_j)$ and the TPR as $\beta_j = \mathbb{E}Y_{ij}\hat{Y}_{ij}/p_j$, where $p_j = \mathbb{P}(Y_j = 1)$. If all the (α_j, β_j) pairs lie on a strictly concave ROC curve $g(\cdot)$, these variables are related by $\beta_j = g(\alpha_j)$. When the patient cases across doctors are drawn from the same population, $p_j = p$ is a constant. It follows from Jensen's inequality that

$$\overline{\beta} = \frac{1}{J}\sum_{j=1}^{J} \beta_j = \frac{1}{J}\sum_{j=1}^{J} g(\alpha_j) < g\left(\frac{1}{J}\sum_{j=1}^{J} \alpha_i\right) = g(\overline{\alpha}),$$

when α_j are not all equal. This important observation appears to have gone largely unnoticed by the literature. Similarly, if each patient i is diagnosed using a point on the human ROC curve $(\alpha_{i,j}, \beta_{i,j})$ and if $\alpha_{i,j}, i = 1, \ldots, n_j$, are not all equal, it is also the case that

$$\beta_j = \frac{1}{n_j}\sum_{i=1}^{n_j} \beta_{i,j} = \frac{1}{n_j}\sum_{i=1}^{n_j} g(\alpha_{i,j}) < g\left(\frac{1}{n_j}\sum_{i=1}^{n_j} \alpha_{i,j}\right) = g(\alpha_j).$$

Online Lemma A.2 provides a set of primitive conditions under which the aggregate human decision maker FPR/TPR lies strictly below the machine ROC curve.

In general, without the restrictions in Assumptions 1 and 2, the decision maker j may use a model $\hat{y} = \mathbb{1}(q_j(x,u) > c_j(x,u))$. For decision maker j minimizing expected loss of misclassification conditional on the feature x and u, consider the following loss matrix:

Loss matrix	$\hat{y} = 0$	$\hat{y} = 1$
$y = 0$	0	$c_{01}(x,u)$
$y = 1$	$c_{10}(x,u)$	0,

where the misclassification losses c_{01} and c_{10} are now dependent on x and u. Equivalently, decision maker j also minimizes the ex ante expected loss:

$$\mathbb{E}[c_{10}(X,U)q_j(X,U)(1-\hat{Y})+c_{01}(X,U)(1-q_j(X,U))\hat{Y}].$$

The optimal decision rule is

$$\mathbb{1}(q_j(x,u)>\frac{c_{01}(x,u)}{c_{10}(x,u)+c_{01}(x,u)}=:c_j(x,u)).$$

Consequently, for (x_1,u_1), (x_2,u_2) sharing the same propensity score $q_j(x_1,u_1)=q_j(x_2,u_2)$, the choice of the decision maker can be different because of cost differences: $c_j(x_1,u_1)\neq c_j(x_2,u_2)$. Such a decision rule is considered in Elliott and Lieli (2013). Even if all doctors have full information-processing capacity so that $q_j(x,u)=p(x,u)$ for a correctly specified propensity function $p(x,u)$, an arbitrary FPR/TPR pair below the machine ROC curve can still be rationalized by some decision rule $\hat{y}=\mathbb{1}(p(x,u)>c_j(x,u))$ corresponding to a reverse-engineered cost function $c_j(x,u)$.

To illustrate using a synthetic example without u, let x be uniformly distributed on $(0,1)$ and $p(x)=x$. Let the cutoff threshold function be

$$c(x)=\begin{cases}0, & x<1/4;\\ 2(x-1/4), & 1/4<x<3/4;\\ 1, & x>3/4.\end{cases}$$

Then, $\hat{y}=\mathbb{1}\left(x<\frac{1}{2}\right)$ and (TPR, FPR) $=(1/8,3/8)$, which lies strictly below the diagonal line. An example when the feature may enter into the cutoff threshold is bank lending, in which $y\in\{0,1\}$ denotes whether a loan is paid back and $\hat{y}\in\{0,1\}$ denotes whether a loan application is approved. The amount of principal and interest are features x that are likely to affect the probability $p(x)$ of defaulting by the borrower but that are also likely to affect the cost perception $c(x)$ by the lender. Assumption 1 essentially excludes such difficulty of interpreting an FPR/TPR pair arising from incentive heterogeneity.

To understand whether the cost perception of doctors might vary systematically with the risk level of their patients, we use the FPR rank for each doctor to measure the conservativeness of the doctor's diagnoses. Doctors may be more willing to tolerate a higher FPR for higher risk patients when they perceive riskier patients as less healthy and more likely to incur a higher cost of underdiagnosis. We then estimate the correlation between the FPR ranks and a pregnancy risk–level indicator. In particular, we regress $q_{i,j}$, the predicted abnormal birth rate for patients diagnosed by doctor j from a random forest algorithm described in Section 3 against r_j, the doctor's FPR rank:

$$q_{i,j}=\beta_0+\beta_1 r_j+\epsilon_{i,j}.$$

The doctors' FPR ranks are normalized to $[0,1]$. If the FPR of doctor j is higher than 80% of the doctors, then $r_j=0.8$. The regression results for doctors with no less than 300 diagnoses and for doctors with no less than 500 diagnoses are summarized in Table 1. The regression results suggest that there is no strong evidence that the FPR rank of doctors is associated with the risk level of their patients.

The use of statistical significance as a benchmark for comparing doctors with the machine algorithm is motivated by the difficulty of fixating a social cost function even if we assume that informed human decision makers are fully rational and impose strong functional form assumptions. Designing a social loss function empirically relies on the domain knowledge of the specific decision maker. Our replacement algorithm allows for the choice of the cutoff threshold to be specific to each individual doctor and does not require a homogeneous preference function to be applied to all doctors. The baseline comparison between true positive rates and false positive rates can also potentially mask the severity of individual cases, the identification of which requires a richer data set with more precise measurements on the outcome variable. Whereas this information is not available in our data set, exploring richer data information can be a fruitful future direction of research.

A responsible phase-in protocol based on multiple years of data on performance observations can incorporate long-term health and well-being outcomes into the definition of performance and quality for decision making. Our data are cross-sectional and do not follow doctors or patients over time. When panel data become available in which doctors and patients can be observed over multiple time periods, future work can potentially generalize the current approach to allow for sequential testing and decision making. The challenge of accounting for longer term health outcomes is partially reflected by the current strategy that replaces a doctor's diagnosis by the machine algorithm only when there is overwhelming statistical evidence suggesting the superiority of the machine decision.

3. Data and Algorithm Description

Our data came from the NFPC project. Starting from 2010, NFPC offers free health checkups for couples planning to conceive 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 examinations and clinical tests, disease and medication history, pregnancy history, and lifestyle and environmental information of both wives and husbands. The data also include the birth outcome, which is denoted as normal ($y=0$) or abnormal ($y=1$). In addition, the data set records doctors' IDs and diagnoses of pregnancy risk. The doctors' diagnoses are graded according to four levels: zero for normal, one for high-risk wife, two for high-risk

Table 1. Predicted Abnormal Birth Probabilities and FPR Rank of Doctors

	Linear term only		With nonlinear term			
	(1)	(2)	(3)	(4)	(5)	(6)
FPR rank	0.003	−0.000	−0.105	0.120	−0.072	0.156
	(0.20)	(−0.01)	(−1.71)	(0.86)	(−0.95)	(0.91)
$Rank^2$			0.111	−0.456	0.072	−0.493
			(1.69)	(−1.32)	(0.89)	(−1.16)
$Rank^3$				0.385		0.380
				(1.61)		(1.28)
Constant	0.206	0.205	0.224	0.205	0.216	0.197
	(21.79)	(18.28)	(17.95)	(14.33)	(14.44)	(11.33)
Number of doctors	584	367	584	584	367	367
R^2	0.0001	0.0000	0.0059	0.0107	0.0025	0.0071

Notes. The table shows regression results for the relation between doctors' preferences and their patients' risk levels. In the baseline linear regression, the dependent variable is the patient's predicted risk by a random forest algorithm. The explanatory variable is the FPR rank of the patient's doctor. Regression results, including quadratic and cubic terms, are reported in columns (3)–(6) in the table. There are 584 doctors diagnosing at least 300 patient cases, and 367 doctors diagnosing at least 500 patient cases. Standard errors in parentheses are clustered by doctors.

husband, and three for high-risk wife and husband. In this paper, we group the doctor's diagnoses into two categories, in which a grade of zero corresponds to a normal pregnancy and any higher grade corresponds to a diagnosis of risky pregnancy.

Our original data set includes 3,330,304 couples that have pregnancy outcomes between January 1, 2014, and December 31, 2015. We exclude duplicate features and samples for which either doctors' diagnoses or more than 50% of feature values are missing. We were left with 168 features of 1,137,010 couples who were diagnosed by 28,716 doctors. Of these observations, 61,184 couples (5.38%) had abnormal birth outcomes.

A basic measurement of the prediction quality of a binary classifier is accuracy, which is the proportion of correct predictions. However, accuracy itself is inadequate in many applications. As the adverse birth rate is about 5%, a naive classifier that categorizes all cases as low risk would achieve an accuracy of nearly 95%. This is clearly controversial. The doctors' overall accuracy is 73.63%, and 24.04% of pregnancies are diagnosed as high risk. Doctors' aggregate false positive rate and true positive rate are 0.2379 and 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 in order to achieve a higher TPR.

We focus on doctors who diagnosed more than 300 patients. There are 584 such doctors corresponding to 584,181 patient cases. Using a ratio of 7:3 and stratification methods, we randomly split our sample into a classification set and a performance set. We use the classification set to train a machine learning algorithm, and then, we categorize a doctor as either "replaced" or "retained" according to our statistical procedure. The performance set is used to compute the FPR/TPR of the combined diagnoses. The data cases are stratified into the four classes based on the actual birth outcomes and the doctors' diagnoses: true positive cases, true negative cases, false negative cases, and false positive cases. The random splitting process is performed through each class of data with the same ratio of 7:3. Subsequently, the data in each class are merged to form the classification and performance sets. In total, there are 408,661 cases in the classification set and 175,520 cases in the performance set.

We further split the classification set into two subsets. The first subset is the training set used to estimate the parameters in a machine learning algorithm. The second subset is the validation set used to evaluate the performance of the algorithm and obtain the machine ROC curve. Splitting of the classification set into the training and validation subsets is done using a ratio of 4:3 and the same stratification method described above. There are 233,435 cases in the training set and 175,226 cases in the validation set.

We experimented with several algorithms. We begin with decision trees, which are widely used in many machine learning applications. However, a single-tree method usually has high variance despite its low bias and tends to overfit by generating results with substantial variation even when a small amount of noise is present in the data. In contrast, RF is a commonly used ensemble learning algorithm proposed by Breiman (2001) and overcomes these problems by constructing a collection of decision trees that are trained using different feature subspaces and bootstrapped samples. The predictions of each tree are aggregated to make predictions. Online Appendix A.2 provides more details about the random forest algorithm.[3]

Remark 1. The classification of doctors' diagnoses by both husband and wife raises a question regarding the validity of Assumption 1 that rules out intraindividual incentive heterogeneity if the relative costs of false positives to false negatives differ between when a single person is deemed risky and when both are deemed risky. A multiclass classification model with

four outcome labels differentiating abnormal pregnancy outcomes because of high risks in wife, husband, and both spouses can allow for heterogeneous relative costs of false positives to false negatives across these cases. However, in our data set, the outcome label is only binary indicating whether the birth is classified as normal or abnormal. The lack of the corresponding four-way outcome label makes it difficult to compute an accurate TPR and FPR for each of the diagnosis classes.

As an illustration, we empirically calculate the FPR/TPR using the binary outcome label for each of the four diagnosis cases and report the aggregate summary in Table 2.

On the one hand, in panels A and B of Table 2, there are some differences between the TPRs and FPRs among the three diagnosing classes of high-risk wife, husband, and both spouses. However, without the corresponding outcome level, it is difficult to interpret these ratios because the total number of abnormal births used in calculating the denominators aggregates the three unobserved latent outcome labels corresponding to high-risk wife, husband, and both spouses.

On the other hand, panel C of Table 2 reports the precision for each of the three diagnosis classes (high-risk wife, husband, and couple). Precision is defined as the fraction of abnormal birth for each of the diagnosis classes. For example, wife (husband) precision represents the fraction of couples with a high-risk wife (husband) diagnosis who have an abnormal birth outcome. Couple precision is similarly defined. In panel C of Table 2, the precision parameters for different diagnosis classes are close to each other. The precision parameters, despite being an imperfect measurement, provide a more accurate indication suggesting that the relative costs of false positives to false negatives are not different between cases when a single person is deemed risky and when both are.

4. Empirical Results for AI and Doctor Decision Making

This section presents the empirical results from the heuristic frequentist approach described in Section 2.3 and the Bayesian approach described in Section 2.4. Doctors who are not replaced by the machine algorithm are called retained doctors hereafter.

4.1. Results of the Heuristic Frequentist Approach

The frequentist approach identifies a dominating segment of the machine ROC for machine algorithm classification. The quality of machine algorithm classification is evaluated by averaging over a uniform grid of points on this dominating segment. For each doctor, we obtain the 95% elliptical confidence area of θ_H centered around the sample estimates using an estimate of the asymptotic covariance matrix based on 100 bootstrap samples. We find the largest TPR point $(\alpha_{\beta_{high}}, \beta_{high})$ and the smallest FPR point $(\alpha_{low}, \beta_{low})$ of the confidence set. The reference point of the doctor is then given by $P = (\alpha_{low}, \beta_{high})$ as shown in Figure 3(a), which is used to classify doctors into two groups. When a doctor's reference point P is below the machine ROC curve, corresponding to case 1 in Section 2.3, the doctor is replaced by the algorithm. When a doctor's reference point P is above the machine ROC curve, corresponding to cases 2 or 3 in Section 2.3, either because this doctor is more capable than the algorithm or this doctor's capability cannot be precisely measured because of fewer patient observations and a larger confidence set, the doctor is retained in decision making.

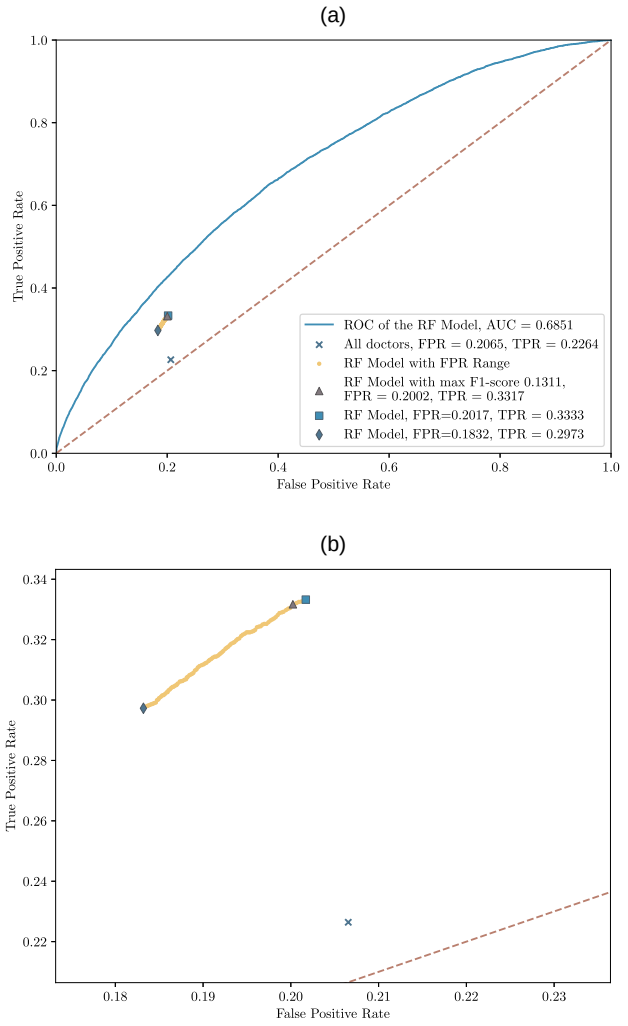
Table 2. Doctor Performance Summary by Multiple Diagnosis Classes

Panel A: FPRs for different diagnosis classes							
Min patient cases	Num negative	FP1	FP2	FP3	FPR1	FPR2	FPR3
300	554,195	51,896	33,266	20,954	0.0934	0.0600	0.0524
500	470,562	42,636	27,362	21,233	0.0906	0.0581	0.0451
Panel B: TPRs for different diagnosis classes							
Min patient cases	Num positive	TP1	TP2	TP3	TPR1	TPR2	TPR3
300	29,986	3,250	1,750	1,941	0.108	0.0584	0.0647
500	24,758	2,589	1,384	1,412	0.104	0.0559	0.0570
Panel C: Precisions across diagnosis classes							
Min patient cases	Wife precision		Husband precision		Couple precision		
300	0.0589		0.0500		0.0626		
500	0.0572		0.0481		0.0624		

Note. In the table, FPR1 denotes FPR where the label is abnormal birth and the diagnosis is wife being high risk; FPR2 and FPR3 are similarly defined for the diagnosis classes of husband being high risk and both spouses being high risk and similarly for TPR1, TPR2, and TPR3.

Among the 584 doctors, 175 doctors are classified by the frequentist approach as replaced by the algorithm. Consequently, patient cases in the performance set are diagnosed by 175 machine doctors (30.0%) and 409 (70.0%) human doctors. For each of the replaced doctors, $N = 100$ points on the dominating segment of the machine ROC curve are chosen corresponding to a uniform grid of the threshold value c for the algorithm to make classification decisions as summarized in Online Appendix A.3. Figure 4 shows the results of this experiment. The aggregate FPR/TPR pair of all doctors on the performance set, represented by the cross, is 0.2065 for FPR and 0.2264 for TPR. The AUC of the random forest on the performance set is 0.6851.

Figure 4. (Color online) Empirical Results of Combining Doctors' and Machine Decisions Using the Heuristic Frequentist Approach



Notes. Doctors' diagnoses $\geqslant 300$. Panel (a) displays the ROC curve of the random forest classifier and the cross representing the FPR/TPR pair of all doctors in the test set. The curved interval shows the aggregate FPR/TPR pairs of combining retained doctors and the machine algorithm using different cutoff threshold choices in the frequentist replacement strategy. Panel (b) provides a magnified view of the curved interval.

The square point uses a decision threshold c corresponding to the upper endpoint of the dominating segment of the machine ROC curve associated with a highest FPR/TPR pair of 0.2017 and 0.3326. Compared with the raw data in which all diagnoses are performed by doctors, the square point increases the TPR by 46.9% and reduces the FPR by 2.3%. The rhombus point corresponds to the lower endpoint of the dominating segment of the machine ROC curve and is associated with a lowest FPR/TPR pair of 0.1836 and 0.2974. Compared with the aggregate doctor diagnoses, the rhombus point increases the TPR by 31.4% and reduces the FPR by 11.1%.

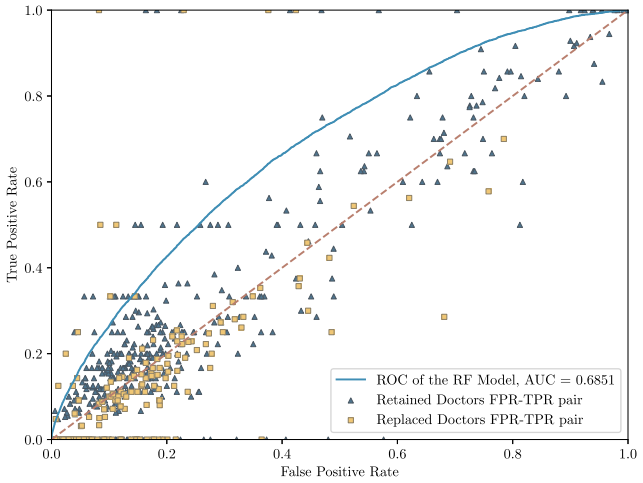
The curved interval in Figure 4 reflects a trade-off between improving the TPR and reducing the FPR. Any point along the interval represents an enhancement over the aggregated doctors' FPR/TPR pair. Furthermore, the triangular point achieves the maximum F1 score[4] of 0.1309 along the dominating segment of the machine ROC curve. Figure 5 is a scatterplot of the FPR/TPR pairs of both replaced and retained doctors against the ROC curve in the performance data set. We see that two adjacent points can be of different replacement status. This can be due to different numbers of patients treated or the use of the test set for calculating the FPR/TPR pairs. An important caveat is that our test depends crucially on the sample size of the numbers of patients treated by each doctor, which introduces an implicit bias that favors experimenting with junior doctors until the null hypothesis that a doctor is not to be replaced is rejected with sufficient evidence from seeing a large number of patient cases.

When the quality of the stock of decision makers may evolve over time, allowing the juniors to learn from making potentially suboptimal decisions for their patients has the potential of increasing the future pool of high-quality seniors. This is likely to be a very delicate issue involving a complex welfare trade-off between current and future generations of patients.

4.2. Results of the Bayesian Approach

The prior distribution $\pi(t)$ is chosen to be a symmetry Dirichlet distribution with the parameters $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ all set to $\gamma = 0.01$. For doctor j's ith patient case, $\hat{Y}_{ij}$ is the doctor's diagnosis and Y_{ij} is the ground-truth label of pregnancy outcome. The likelihood for data is given in (5), where $\hat{Y}_i$, Y_i, and n in (4) are replaced by $\hat{Y}_{ij}$, Y_{ij}, and n_j. The posterior distribution is also a Dirichlet distribution with parameters $(\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \hat{\gamma}_4)$, where $\hat{\gamma}_k = \gamma + n\hat{t}_k$ for $k = 1, 2, 3, 4$. The posterior distribution of $\theta_H = (\alpha_H, \beta_H)$ can be simulated in a two-step procedure. In the first step, R samples of $\{\boldsymbol{t}_r = (t_{1r}, t_{2r}, t_{3r}, t_{4r})\}_{r=1}^R$ are drawn from the posterior Dirichlet distribution with parameters $(\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \hat{\gamma}_4)$. In the second step, we compute $\{\theta_{H,r} = (\alpha_H(\boldsymbol{t}_r), \beta_H(\boldsymbol{t}_r))\}_{r=1}^R$ with the simulated R samples of $\boldsymbol{t}_r, r = 1, \ldots, R$, using (6).

Figure 5. (Color online) Scatterplot of Replaced and Retained Doctors in the Test Data Set Using the Frequentist Approach



Notes. Doctors' diagnoses $\geqslant$ 300. The FPR/TPR pairs in the scatterplot are calculated in the test data set of patient cases for each doctor. It is possible for the FPR/TPR pairs of some of the replaced doctors to lie above the machine ROC curve in the test data set. Two points that are quite close to each other can be of different replacement status because of different numbers of patients treated.

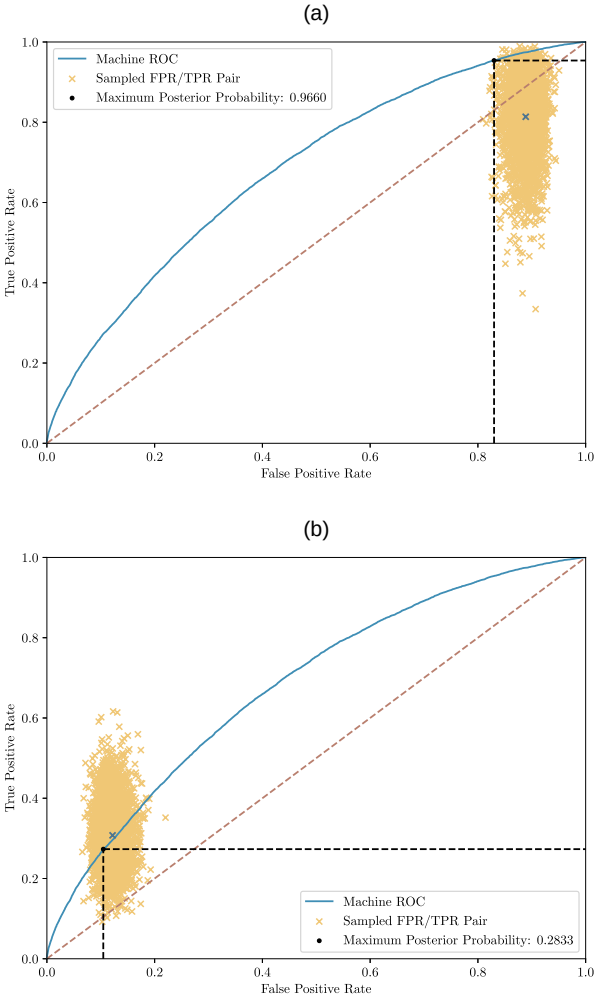
The simulated draws $\theta_{H,r}, r = 1, \ldots, R$ are used to calculate (10). If $\hat{q}_{max} \geqslant 0.95$, we mark the doctor as replaced. The threshold c that corresponds to the point $(\hat{\alpha}_{max}, g(\hat{\alpha}_{max}))$ given by (10) is used as the decision threshold for the algorithm. In Figure 6(a), the maximum coverage rate $\hat{q}_{max} = 0.9657$ at the dot on the machine ROC curve. The doctor is replaced by the algorithm. In Figure 6(b), the maximum coverage rate $\hat{q}_{max} = 0.3179$. There is not sufficient evidence to replace the doctor's decisions with the algorithm's.

Figure 7(a) shows the experiment results for doctors who had diagnosed more than 300 cases. A total of 269 (46.1%) out of 584 doctors are classified as replaced. The remaining 315 (53.9%) are retained human doctors. The combined decisions by the retained human doctors and the algorithm's decisions indicated in the figure by the triangular point, result in an FPR of 0.1856 and a TPR of 0.3319, corresponding to an increase of 46.6% in the TPR and a reduction of 10.1% in the FPR. Similar to Figure 5, Figure 7(b) is a scatterplot of the FPR/TPR pairs of both replaced and retained doctors against the ROC curve in the performance data set.

We experiment with several alternative definitions described in Equations (12)–(15) for the Bayesian posterior expected loss functions in (11). These results are shown in Figure 8, in which panel (b) magnifies a portion of panel (a) containing the replacement paths. Ranking the doctors by their expected posterior losses allows us to replace only the doctors whose expected loss of being substituted by the machine algorithm is relatively low. Figure 8 traces the dynamics of the aggregate FPR/TPR pairs of combining the doctor with the machine decisions when the ratio of replaced doctors increases from 0% to 100%.

In the Figure 8(a) legends, *Bayesian Test* refers to the baseline construction in (8), *Euclidean Distance to the ROC Curve* refers to the loss function in (12), *Complementary Set Boundary Euclidean Distance* refers to the loss function in (13), *Diagonal Line Horizontal and Vertical Decompositions* refer to the loss function in (14), and *Complement Set Boundary Vertical and Horizontal Decompositions* refer to the loss function in (15). No individual loss function demonstrates saliently stronger discriminatory power over the entire replacement path. Furthermore, the incremental benefit going beyond a 50% replacement rate is a small marginal reduction of FPR. In

Figure 6. (Color online) Illustration of Replaced and Retained Doctors by the Bayesian Approach



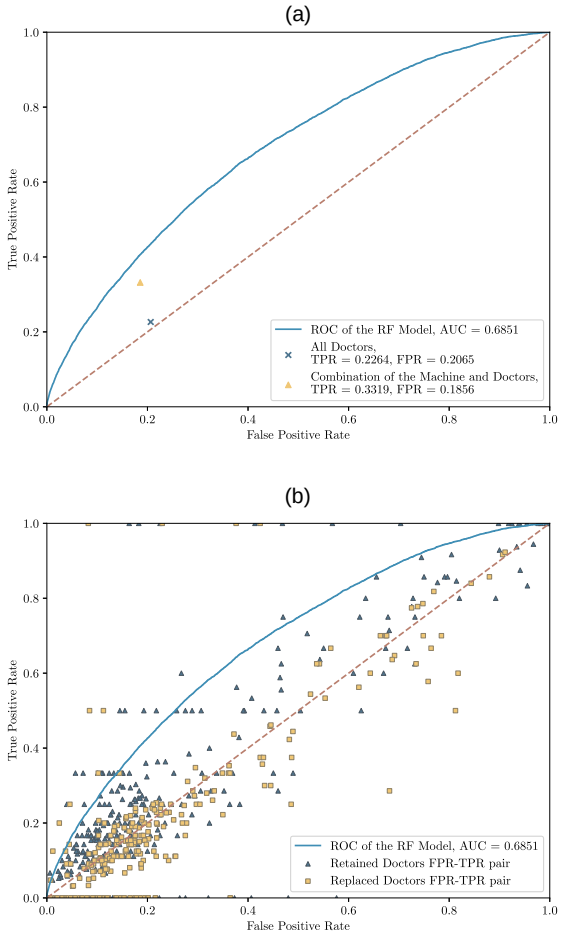
Notes. In these two figures, the crosses are sampled from the projection of the posterior Dirichlet distribution. One thousand draws are made for each panel. On the machine ROC curves, the dots are the points with the largest dominating posterior probability mass. Panel (a) shows a doctor whose posterior probability of being dominated is greater than 0.95. This doctor will be replaced by the machine decision rule corresponding to the dot. Panel (b) illustrates the opposite case in which the doctor will not be replaced.

addition, the baseline 95% credible level Bayesian test implemented using (8) is not dominated by any of the competing methods in the FPR/TPR comparison.

The null hypothesis of both the Bayesian analysis and the frequentist analysis is the status quo of human decision making. A human decision maker is replaced by the machine algorithm only when there is strong evidence from the data favoring the machine algorithm. An alternative conceptual paradigm is to treat the machine algorithm as the status quo null hypothesis, in which the human decision makers are by default replaced by the machine algorithm unless there is strong data evidence validating the superiority of human decision making.

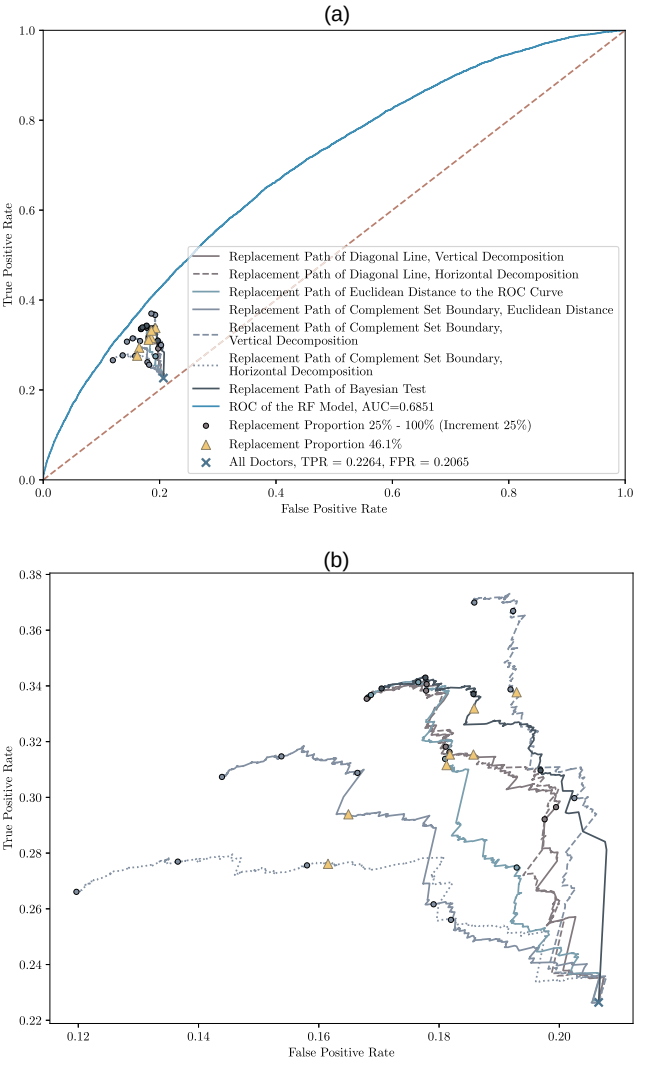
However, when we experiment with the alternative specification of the status quo, all except a few doctors are determined to be replaced by the machine algorithm. In the cross-point labeled as the dominate

Figure 7. (Color online) Empirical Results of the Bayesian Approach



Notes. Doctors' diagnoses $\geqslant$ 300. Panel (a) draws the ROC curve in the test set of the random forest classifier, the FPR/TPR pair of all doctors, and the FPR/TPR pair of doctor/machine combination. The panel (b) scatterplot represents the FPR/TPR pairs of individual doctors such that square points correspond to the replaced doctors and the triangular points nonreplaced doctors.

Figure 8. (Color online) Bayesian Approach with Loss Functions Constructed by Euclidean Distance and Randomized Decomposition

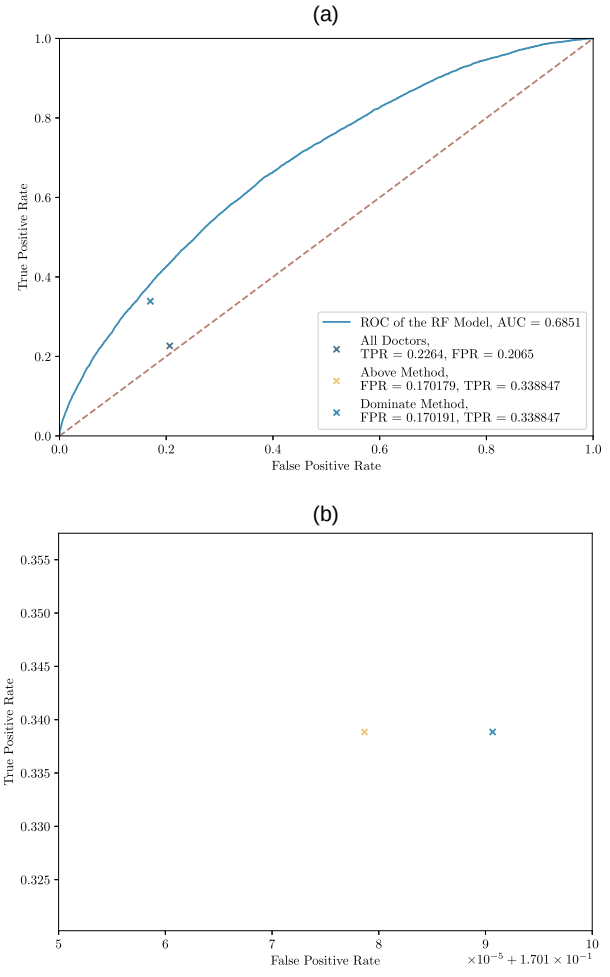


Notes. Doctors' diagnoses $\geqslant$ 300. The replacement paths trace the dynamics of the aggregate FPR/TPR pairs of combining the doctor decisions with the machine decisions when the replacement ratio increases from 0% of decisions entirely by human to 100% of decisions entirely by machine algorithms. Panel (b) magnifies a portion of panel (a) containing the replacement paths.

method in Figure 9, a doctor is retained if there exists a point on the machine ROC curve that is dominated by at least 95% of the posterior distribution of $p(\theta_H|\mathbb{D})$. A mere six doctors are retained in this method. In the cross-point labeled as the above method in Figure 9, a doctor is retained if the posterior probability of $p(\theta_H|\mathbb{D})$ above the machine ROC curve is more than 95%. Only seven doctors are retained using this method. For doctors who are replaced by the machine algorithm, the replacement point on the ROC curve is determined by the baseline Bayesian tests in (10).

At high replacement ratios, the FPR and TPR levels decrease simultaneously across all loss functions. Even

Figure 9. (Color online) Alternative Null Hypothesis of Machine Algorithm: Doctors Are Replaced by Default and Are Retained Only with Strong Evidence



Notes. In the "dominate method" cross point, a doctor is retained if there exists a point on the machine ROC curve that is dominated by at least 95% of the posterior distribution of $p(\theta_H|\mathbb{D})$. A mere six doctors are retained in this method. In the "above method" cross point, a doctor is only retained if the posterior probability of $p(\theta_H|\mathbb{D})$ above the machine ROC curve is more than 95%. Only seven doctors are retained using this method. These two points are almost indistinguishable in panel (a). Their difference is very small as shown in (b).

when all the doctors are substituted by the machine algorithm, the resulting FPR/TPR pair still lies below the machine ROC curve. This is because different threshold cutoff values on the machine ROC curve are employed to replace different doctors and because of the concavity arguments formalized in Online Lemma A.2.

On the one hand, in the current empirical data set, only a small fraction of doctors have FPR/TPR pairs above the machine ROC curve. We have not been able to combine doctors with the machine algorithm to generate an FPR/TPR pair above the machine ROC curve. On the other hand, using a data-generating process that allows for a substantial portion of doctors with FPR/TPR pairs to be above the machine ROC curve, reported in the synthetic data analysis Section 5, we are able to combine the doctors and the machine algorithm to generate a FPR/TPR pair that outperforms even the machine ROC curve.

Instead of deterministically replacing some doctors by the machine algorithm, Online Appendix A.5 reports the results of a randomized replacement experiment in which each replaced doctor accepts the machine decision probabilistically according to a prespecified acceptance rate $\lambda \in [0,1]$. As expected, the overall FPR/TPR pairs vary incrementally from when $\lambda = 0$, corresponding to human decision making, to when $\lambda = 1$, corresponding to deterministic replacement of doctors by the machine algorithm.

Online Appendix A.6 considers a more general form of substitution to replace a subset of patient cases for each doctor. We use patient case covariates to predict the likelihood that the doctor's diagnoses differ from the ground-truth outcome and replace only those patient cases in which the doctor is prone to make mistakes. Whereas the resulting pattern of the replacement paths is quite different, the patient-case specific substitution approach does not show saliently stronger discriminatory power compared with the baseline Bayesian approach.

4.3. Characteristics of Replaced Doctors

Our paper identifies a subset of the doctors to be replaced by the machine algorithms according to FPR/TPR indicators. We create a dummy variable *replaced* that is set to one when a doctor is replaced by the algorithm and is set to zero otherwise and examine the partial correlation coefficient between the *replaced* dummy variable with several confounding factors, including whether they practice in a clinic or hospital.

We regress the *replaced* indicator on three confounding factors. The first factor is county-level per capita gross domestic product (GDP) (in 10,000 Chinese yuans) collected manually from the statistical yearbooks of 2014 across cities and provinces, which is the starting year of our data.[5] The second factor is the logarithm of the total number of patient cases diagnosed by each doctor. The third factor is represented by two dummy variables related to the practice history of each doctor. A doctor in our data set may practice in more than one type of facility. The practice facility can be either a county hospital or a township clinic. Medical facilities in China are typically classified by the administrative level to which they belong. On the one hand, urban residents usually go to a county-level hospital. On the other hand, rural residents oftentimes go to a township-level clinic, which is generally considered to be of lower quality than a county-level hospital. The doctors in our data fall into three categories. Doctors who have practiced in both types of facilities are used as the reference level.

The first dummy variable representing the third factor is set to one for doctors who have practiced only in a county hospital. The second dummy variable representing the third factor is set to one for doctors who have practiced only in a township clinic. There are additional variables in the data set that are only available at a more aggregated city or province level, such as birth rate and the average number of medical staff. We do not use these variables because, given how administrative units are divided in China, a municipal unit can include many county-level units with heterogeneous economic conditions.

The data summary statistics are reported in panel A of Table 3. As shown in panel A of Table 3, doctors who have practiced at both a county hospital and a township clinic have seen the largest number of patient cases and have a high replacement ratio. Doctors who have practiced only at a township clinic have seen a smaller number of patient cases compared with doctors who have practiced only at a county hospital, but doctors who have practiced only at a township clinic have a higher replacement ratio than doctors who have practiced only at a county hospital have.

Panel B in Table 3 reports a logistic regression of the replacement indicator on county level per capita GDP and the logarithm of the number of patients seen by the doctor. Column (2) adds an indicator of whether the doctor practices only in a township clinic and an indicator of whether the doctor practices only in a county hospital. The reference level at which both indicators are zero corresponds to doctors who have practiced in both a township clinic and a county hospital. Province fixed effects are included in the logistic regression. We use cluster robust standard errors in order to correct for spatial autocorrelation.

Table 3. Characteristics of Replaced Doctors

Panel A: Summary statistics by doctor type

	Num	Log num	GDP	Replacement ratio
County & town	276	6.797	3.956	0.522
		(0.73)	(2.94)	
County only	161	6.515	3.675	0.398
		(0.65)	(2.27)	
Town only	43	6.409	3.038	0.512
		(0.58)	(2.26)	

Panel B: Logistic regression			Panel C: TPR kernel regression	
	(1)	(2)	(3)	(4)
Log num	0.896	0.885	0.000251	0.000117
	(2.85)	(3.20)	(0.04)	(0.02)
GDP percap	−0.0114	−0.0124		0.000569
	(−0.37)	(−0.41)		(0.28)
Count only		−0.367	0.0172	0.0181
		(−1.99)	(2.04)	(1.99)
Town only		0.359	−0.0273	−0.0235
		(0.86)	(−1.73)	(−1.63)
FPR			0.958	0.954
			(44.02)	(51.70)
Observations	457	457	480	469

Notes. Panel A shows the summary statistics of doctors grouped by type of facility. Panel B shows regression results of a logistic regression with province-level fixed effect. Cluster robust *t*-statistics are reported in parentheses. A more direct analysis of doctor performance is provided by a kernel regression of TPR on FPR and other covariates reported in Panel C. As in panel B, *t*-statistics are given in parentheses.

The logistic coefficient estimates do not contradict our a priori reasoning. Whereas higher per capita GDP is associated with a lower likelihood of replacement, the corresponding coefficients are small and lack statistical significance. The coefficients of the log number of patients are positive and statistically significant, indicating the higher distinguishing power of the Bayesian test when a doctor sees more patients. The FPR/TPR pairs of doctors with more patient cases can be estimated more precisely. Consequently, it is more likely for these doctors whose FPR/TPR pairs are below the machine ROC curve to be identified as replaced by our statistical procedure. This finding is consistent with the lack of evidence contradicting Assumption 2 that doctors accumulate experience from additional patient cases.

Doctors practicing only at township clinics have higher replacement probability than reference level doctors practicing at both township clinics and county hospitals, who, in turn, have higher replacement probability than doctors practicing only at county hospitals. Anecdotal evidence in China suggests that doctors with better medical school records are more likely to be assigned directly to county-level hospitals than to have to work their way up from township clinics into county hospitals.

To distinguish an analysis of the predictors of performance of doctors from the predictors of tests concluding whether they are to be replaced by the machine algorithm, we implement a kernel regression of TPR on FPR and a number of confounding features, including county-level per capita GDP, the logarithm of the number of patients seen by the doctor, an indicator of whether the doctor practices only in a township clinic, and an indicator of whether the doctor practices only in a county hospital.[6] Among doctors achieving the same FPR, those achieving a higher TPR can be considered to perform better. A confounding variable with a positive coefficient enhances the TPR after controlling for FPR and is interpreted as a variable that positively predicts performance.

Panel C in Table 3 reports the results from this regression. The regression coefficients have signs that are consistent with a priori reasoning, but they are also mostly insignificant except for the coefficients on the indicator of whether the doctor practices only in a county hospital. For a given FPR, a higher TPR is positively associated with higher county-level per capita GDP, a larger number of patients seen by the doctor, and the indicator

that the doctor practices only in a county hospital but is negatively associated with the indicator that the doctor practices only in a township clinic. The statistical insignificance of the coefficients on the logarithm of the number of treated patients in panel C of Table 3 and their significance in panel B of Table 3 provide further evidence that the replacement indicator is largely driven by the difference in the sample size. A larger sample size makes it easier to reject the null hypothesis that the machine algorithm does not outperform the doctor.

5. Synthetic Data Analysis

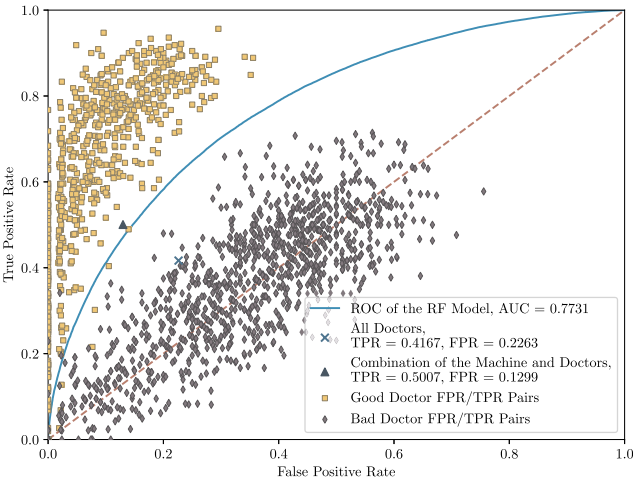
5.1. An Example of Human–Algorithm Complementarity

In our empirical data set, only a small fraction of doctors lie above the machine ROC curve. Combining the machine algorithm and the doctor decisions results in aggregate FPR/TPR pairs that are still below the machine ROC curve. In addition, Jensen's inequality (elaborated in Section 2.5) likely contributes to the combined aggregate FPR/TPR pair lying below the machine ROC curve. Even if we replace every doctor by the machine algorithm, unless the same point on the machine ROC curve is used to replace all doctors, the combined aggregate FPR/TPR pair is necessarily below the machine ROC curve. The endpoints of the replacement paths in Figure 8 correspond to replacing all doctors by the machine algorithm. The algorithms that we implemented use different points on the machine ROC curve to replace doctors. These points also differ across different doctors. In order for the combined aggregate FPR/TPR pair to be above the machine ROC curve, it is necessary that a sizable portion of the original doctor diagnoses lie above the machine ROC curve.

Whereas not the case in the NFPC data set, other data sets might exist in which a significant portion of human decision makers correctly use private information not available to the machine algorithm. Figure 10 reports the results from a synthetic data analysis in which the combined aggregate FPR/TPR pair between doctors and the machine algorithm lies above the machine ROC curve.

The data-generating process of the simulation in Figure 10 is as follows. The input features are three-dimensional: $x_{i,j,1}$, $x_{i,j,2}$, and $u_{i,j}$. On the one hand, the first two features, denoted $x_{i,j,1}$ and $x_{i,j,2}$, are assumed to be observable by both the doctors and the machine algorithm. On the other hand, the last feature, denoted $u_{i,j}$, is only observable to the doctors and is not used by the machine algorithm. The three features are generated from independent normal distributions with prespecified means and variances: $X_{i,j,1} \sim N(0, 1.95)$, $X_{i,j,2} \sim N(0, 0.25)$, and $U_{i,j} \sim N(0, 2)$. The ground-truth variable

Figure 10. (Color online) Synthetic Data for Which Combined FPR/TPR Is Above the Machine ROC Curve



Notes. In the figure, we simulate 750 doctors with better information than the machine algorithm and full information processing capabilities. These doctors are represented as the square FPR/TPR points. Meanwhile, we simulate 1,250 doctors with the same information but use a misspecified prediction model in their decisions. These doctors are plotted as the rhombus FPR/TPR points. Then, we apply our Bayesian test approach to identify and replace the less capable doctors in the simulation. As shown in the figure, the doctor–algorithm combined aggregate FPR/TPR point lies above the machine ROC curve, an example of complementarity.

$Y_{i,j} \in \{0, 1\}$ is generated by

$$Y_{i,j} = \mathbb{1}(p(X_{i,j,1}, X_{i,j,2}, U_{i,j}) > \delta_{i,j}) \quad \text{where}$$

$$p(X_{i,j,1}, X_{i,j,2}, U_{i,j}) = \frac{e^{X_{i,j,1}+X_{i,j,2}+U_{i,j}}}{1 + e^{X_{i,j,1}+X_{i,j,2}+U_{i,j}}}$$

and where $\delta_{i,j}$ is uniformly distributed between zero and one. Using this specification of the data-generating process, we simulated 600,000 patient cases and allocate them evenly to 2,000 doctors. Each doctor is assigned 300 simulated patient cases.

Doctors are partitioned into two groups according to their information-processing capacities. The first group, consisting of 750 more capable doctors, use the correct full-information propensity score $p(x_{i,j,1}, x_{i,j,2}, u_{i,j})$ with different incentives to diagnose patient cases. Their decision rule is given by, using C_j that is uniformly distributed between 0.4 and 1 across doctors,

$$\hat{Y}_{i,j} = \mathbb{1}(p(X_{i,j,1}, X_{i,j,2}, U_{i,j}) > C_j).$$

The second group of doctors, consisting of 1,250 less capable doctors, use a misspecified propensity score to diagnose patients. Their decision rule is

$$\hat{Y}_{i,j} = \mathbb{1}(q(X_{i,j,1}, X_{i,j,2}, U_{i,j}) > C_j) \quad \text{where}$$

$$q(X_{i,j,1}, X_{i,j,2}, U_{i,j}) = \frac{e^{-X_{i,j,1}+X_{i,j,2}+U_{i,j}}}{1 + e^{-X_{i,j,1}+X_{i,j,2}+U_{i,j}}}$$

is a misspecified propensity score function.

We divide the 600,000 simulated observations into a training subset consisting of 240,000 observations, a validation subset consisting of 180,000 observations, and a test subset consisting of the remaining 180,000 observations. We estimate a random forest model using only the training subset and the first two features, $x_{i,j,1}$ and $x_{i,j,2}$, and use the baseline Bayesian test in Equation (10) to combine doctors with the machine algorithm. The triangular point in Figure 10 represents the combination of doctors and the machine algorithm. It lies above the ROC curve of the random forest model and also strictly dominates the cross of the aggregate FPR/TPR pair of all doctors. In summary, the synthetic data analysis represented by Figure 10 illustrates the scientific principle of complementarity in human algorithm collaborations. See, for example, Bansal et al. (2021).[7]

5.2. Aggregating Information by Predicted Doctor Models

Our data set contains doctor diagnosis information in addition to the eventual pregnancy outcomes. In this section, we discuss whether estimating a model that predicts the label of doctors' diagnoses can assist in improving the FPR/TPR trade-off of medical decisions. A model of predicted doctors can be intuitively interpreted as doctors' collective wisdom. The experience of senior doctors can provide guidance to junior doctors with limited clinical exposure.

In the NFPC data set, a model of predicted doctors does not appear to improve on the aggregate FPR/TPR pair. However, in several of the following synthetic data examples, a predicted doctor model improves upon the diagnoses of individual doctors. Each doctor's diagnosis may encode information heterogeneity or incentive heterogeneity or both. Consider a model of incentive heterogeneity, in which doctors' decision rules are given by $\hat{y}_i = \mathbb{1}(p(x_i) > u_i)$ such that $p(x_i)$ is the correctly specified propensity score and u_i is an independent distribution with a strictly monotonic cumulative distribution function $G(\cdot)$. On the one hand, the aggregate doctor FPR/TPR lies below the optimal ROC curve. On the other hand, the predicted doctor model, given by $\mathbb{E}[\mathbb{1}(p(X_i) > U_i)|X_i] = \mathbb{P}(U_i \leqslant p(X_i)) = G(p(X_i))$, is a monotonic transformation of the true propensity score function and recovers the optimal ROC curve.

We next focus on several information models without incentive heterogeneity. Consider a baseline example in which doctors have access to private information u_i in addition to the publicly observed features x_i and are aware of the correct full-information propensity score model $p(x_i, u_i)$. If we could directly observe doctors' probabilistic assessment $p(x_i, u_i)$ and use it as the label to train a machine learning model, then the resulting model necessarily coincides with the ground-truth machine-learned propensity score model:

$$\mathbb{E}[p(X_i, U_i)|X_i] = \mathbb{E}[\mathbb{E}[Y_i|X_i, U_i]|X_i] = \mathbb{E}[Y_i|X_i] = p(X_i).$$

However, in reality, we rarely solicit doctors' probabilistic assessment of the likelihood of an illness. Instead, we only observe the binary diagnosis by the doctors, which corresponds to observing $\hat{y}_i = \mathbb{1}(p(x_i, u_i) > c_0)$ for some threshold value c_0. When we use $\hat{y}_i$ as the label to train a machine learning model, the propensity score of the resulting predicted doctor model is

$$q(X_i) = \mathbb{E}[\mathbb{1}(p(X_i, U_i) > c_0)|X_i].$$

In general, $q(X_i)$ is different from $p(X_i)$. The ROC curve implied by $q(X_i)$ coincides with $p(X_i)$ only when $q(X_i)$ is an increasing transformation of $p(X_i)$: $q(X_i) = \Lambda(p(X_i))$, where $\Lambda(\cdot)$ is a strictly increasing function. There is no guarantee that this is necessarily the case.

Consider now three scenarios in which doctors employ misspecified models to process private information. In the first scenario, doctors process the full information (X_i, U_i) using an incorrect propensity score model $q(X_i, U_i)$, but the ROC curve implied by the predicted doctor model $q(X_i)$ still coincides with the ground-truth machine ROC curve. Figure 11(a) illustrates this scenario using the following data-generating process. The ground-truth propensity score model is $p(X_i) = \exp(X_i)/(1 + \exp(X_i))$. The doctors use a misspecified propensity model of $q(X_i, U_i) = \exp(X_i + U_i)/(1 + \exp(X_i + U_i))$. Both X_i and U_i are uniformly and independently distributed: $X_i \sim U(-1, 1)$ and $U_i \sim U(-2, 2)$. The "X_i only" ROC curve is generated by $p(X_i)$; the "X_i and U_i" ROC curve is generated by $q(X_i, U_i)$. The "Predicted Doctors on X" ROC curve is generated by $q(X_i) \equiv \mathbb{E}[\mathbb{1}(q(X_i, U_i) > c_0)|X_i]$.

For any given cutoff threshold c_0, the predicted doctor model $q(X_i)$ is an increasing function of X_i as is $p(X_i)$. In this particular case, both $p(X_i)$ and $q(X_i)$ correspond to the same ROC curve. This curve lies above the ROC curve generated by $q(X_i, U_i)$ and the aggregate doctor FPR/TPR pair ($c_0 = 0.7$). Collective wisdom improves upon the misspecified model employed by individual doctors and results in higher diagnosing quality as represented by the ROC curves.

The second scenario, pictured in Figure 11(b), is generated by a two-dimensional observable feature model. The ground-truth propensity score model is

$$p(X_{i1}, X_{i2}) = \exp(X_{i1} + X_{i2})/(1 + \exp(X_{i1} + X_{i2})).$$

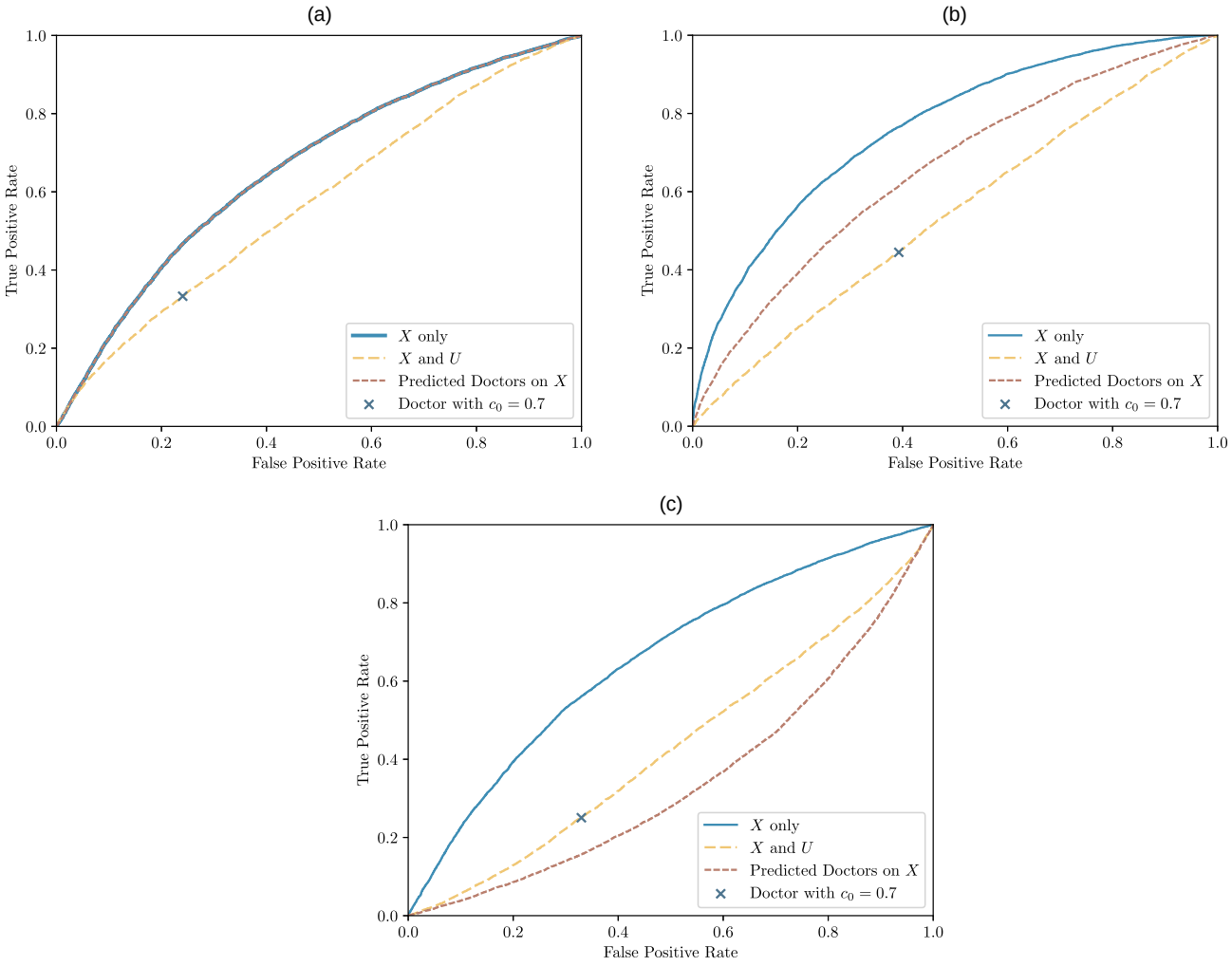
The doctors' (misspecified) information model is

$$q(X_{i1}, X_{i2}, U_i) = \exp(X_{i1} - X_{i2} + U_i) / (1 + \exp(X_{i1} - X_{i2} + U_i)).$$

Here X_{i1}, X_{i2}, and U_i are independently distributed: $X_{i1} \sim N(0, 1), X_{i2} \sim N(0, 0.5), U_i \sim N(0, 4)$.

In this scenario, doctors also process the full information (X_{i1}, X_{i2}, U_i) using an incorrect propensity score

Figure 11. (Color online) Three Cases of Learning from Doctors



Notes. Instead of predicting ground-truth outcome, this figure illustrates three synthetic data examples in which machine algorithms are applied to predict doctor's diagnoses. Panel (a) shows a scenario in which the machine algorithm improves upon the individual doctor diagnoses and generates the same ROC curve as a correctly specified model. Panel (b) shows a scenario in which aggregating information from doctor diagnoses is beneficial but not sufficient to recover the optimal ROC curve. The last case is shown in Panel (c) in which aggregating information from doctor diagnoses results in a ROC curve inferior to that of the misspecified prediction model.

model. Whereas the ROC curve corresponding to the predicted doctor model $q(X_{i1}, X_{i2}) \equiv \mathbb{E}[\mathbb{1}(q(X_{i1}, X_{i2}, U_i) > c_0)|X_{i1}, X_{i2}]$ lies below the optimal ROC curve corresponding to the correct propensity score model $p(X_{i1}, X_{i2})$, it still lies above the ROC curve corresponding to the individual doctor misspecified information model $q(X_{i1}, X_{i2}, U_i)$ and the aggregate doctor FPR/TPR pair. Collective wisdom in this case leads to limited improvement over the misspecified information model.

The third scenario is pictured in Figure 11(c). The ground-truth propensity score model is $p(X_i) = \exp(X_i)/(1 + \exp(X_i))$, but the doctors' incorrect information model is

$$q(X_i, U_i) = \exp(-X_i + U_i)/(1 + \exp(-X_i + U_i)),$$

where X_i and U_i are distributed as in the first scenario depicted in Figure 11(a). In this model, doctors' information-processing capacity is very limited. The predicted doctor model $q(X_i) = \mathbb{E}[\mathbb{1}(p(X_i, U_i) > c_0)|X_i]$ exacerbates the misspecification of the doctors' private information propensity score model and leads to an ROC curve that lies strictly below the ROC curve corresponding to the individual doctor's misspecified propensity score model $q(X_i, U_i)$.

The analysis in Figure 11 shows that, depending on the extent of misspecification of the doctor's information model, the predicted doctor model that aggregates the information among doctors using the subset of observable features available to the machine learning algorithm might either reduce or exacerbate the misspecification in the propensity score model. The resulting ROC curve of the predicted doctor model

might lie above or below the ROC curve of the doctors' information model.

5.3. Replacement Strategies Initiated by Doctors

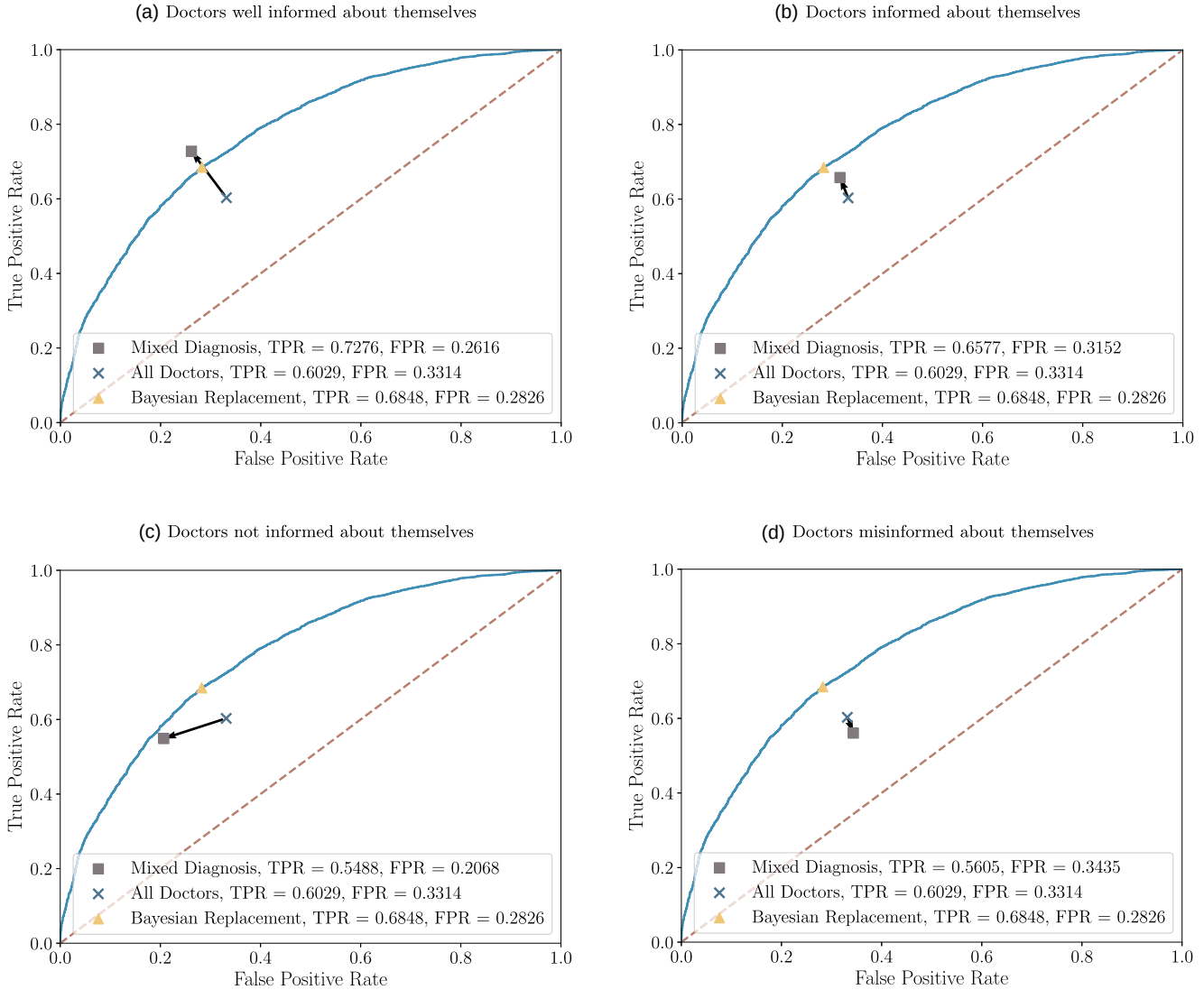
For a given decision maker, recognizing that machine algorithms are more likely to be applicable in some candidate cases than in others is important for a replacement model to achieve human and machine complementarity. Formalizing such a general empirical replacement model is difficult because of the unknown diagnosing model employed by doctors in the data and because of our lack of knowledge of doctors' preferences for the machine algorithm. To illustrate these difficulties, we experiment with a set of simulations in which doctors decide when to employ the machine learning algorithm.

In the simulation setup, each patient is associated with three independently generated, normally distributed features X_1, X_2, U, where $X_1 \sim N(0, 0.5)$, $X_2 \sim N(0, 1.5)$, and $U \sim N(0, 1)$. The probability of each patient being ill is a logit function of the features

$$p(X_1, X_2, U) = e^{X_1+X_2+U}/(1 + e^{X_1+X_2+U}).$$

The labels are generated according to the probability model $p(X_1, X_2, U)$ for 30,000 patient cases. The diagnosis of doctors is modeled as a mixture between two decision rules. The first rule employs the correct probability model $p(X_1, X_2, U)$. The second rule replaces the probability model $p(x)$ by an uninformative uniformly

Figure 12. (Color online) Replacement Result Under Different Replacement Rules



Notes. A flexible replacing model allowing doctors to choose whether to apply the algorithm's decision has the potential of achieving human–machine complementarity, whereas in this figure, we show that the performance of this strategy highly depends on doctors' choices. If the doctors do not have a clear recognition of their information weakness, this strategy may not perform well.

distributed random variable $r \sim U(0,1)$. The second rule applies to a situation in which the occasional lack of medical knowledge by doctors results in randomized diagnosis. Specifically, we simulate doctor diagnosis $\hat{Y}$ by

$$\hat{Y} = \begin{cases} \mathbb{1}(e^{X_1+X_2+U}/(1+e^{X_1+X_2+U}) > c_d), & \text{if } X_1 \geqslant 0, \\ r > c_d, & \text{otherwise.} \end{cases}$$

In the above, $c_d = 0.5$ is the decision threshold for doctor diagnosis. The simulated data set is divided into training, validation, and test subsets by the ratio of 4:3:3. They are used to estimate model parameters, compute the ROC curve, and calculate FPR/TPR pairs, respectively. We first train a random forest model using features X_1 and X_2, in which U is interpreted as unobserved heterogeneity that is known to the doctors but not to the machine algorithm.

We simulate four replacement rules. They are illustrated in Figure 12. In the first rule, shown in Figure 12(a), doctors know when their medical knowledge is inadequate (e.g., when $X_1 < 0$) and choose to defer to the machine algorithm in such a situation. This method achieves complementarity between the human doctor and the machine algorithm. It produces an FPR/TPR pair that is better than the Bayesian approach in which either all or none of the patients for each doctor are replaced. In the second rule, shown in Figure 12(b), doctors have partial but incomplete knowledge of when their medical knowledge is inadequate. Doctors defer to the machine algorithm when $X_1 < 0.5$. The second rule improves over the doctors but does not outperform the Bayesian approach. Figure 12(c) use a replacement rule of $X_2 \leqslant -0.5$. In this case, the doctor does not know when the doctor lacks medical knowledge, and the replacement reduces both FPR and TPR. In Figure 12(d), doctors mistakenly defer to the machine algorithm when they do have medical knowledge (e.g., when $X_1 \geqslant 0$) but make their own decision when they are uninformed (e.g., when $X_1 < 0$). The resulting FPR/TPR pair is strictly dominated by the original doctor diagnosis.

6. Conclusion

In this paper, we propose statistical tests for replacing human decision makers with algorithms. Decision rules that robustly dominate diagnoses made by humans can be determined based on the machine ROC curves. We propose both a heuristic frequentist approach and an alternative Bayesian posterior loss function approach. By replacing a subset of decision makers with algorithms, we can achieve a higher quality decision-making outcome. We experiment with the two approaches using a data set of prepregnancy checkups and find that the replacement of a subset of doctors using the Bayesian approach significantly improves the overall performance of risky pregnancy detection. Furthermore, our results indicate that doctors who practice only at township clinics are more likely to be replaced, whereas doctors who practice only at county hospitals are less likely to be replaced.

Our analysis shows that both stable preference and decision-making quality are important for comparing machine algorithms with human decision makers. Whereas allowing the juniors to make potentially suboptimal decisions for their patients and learn from the results has the potential of increasing the pool of high-quality seniors, this is a very delicate issue involving a complex welfare trade-off between current and future generations of patients. In panel data sets with rich dynamic information, human decision makers can engage in information updating and peer learning. Our current data set is only cross-sectional. The eventual pregnancy outcomes are typically not available to doctors when they diagnose consecutive patient cases. Juniors might also benefit more generally from off-line learning without having to make substantial and high-stakes decisions.

Acknowledgments

The authors thank the editors and four anonymous referees, Brendan Beare, Bo Honore, Xiaohong Chen, Ron Gallant, Bruce Hansen, Peter Hansen, Bentley MacLeod, Jack Porter, Adam Rosen, Andres Santos, George Tauchen, Valentin Verdier, and participants at various seminars and conferences for insightful comments. The authors also thank Yichuan Zhang, Xin Lin, and Zhonghao Huang for excellent research assistance. This study was approved by the National Health Commission. Informed consents were obtained from all the NFPC participants.

Endnotes

[1] See, for example, Long et al. (2017), Kermany et al. (2018), De Fauw et al. (2018), Rajpurkar et al. (2017), Hannun et al. (2019), Esteva et al. (2017), Han et al. (2020), Daneshjou et al. (2022), Ardila et al. (2019), Liang et al. (2019), McKinney et al. (2020), Peng et al. (2021), Uhm et al. (2021), Tian et al. (2024), Berk (2017), and Kleinberg et al. (2018).

[2] We are grateful to the associate editor for pointing out this important clarification.

[3] We choose the following parameter settings. The number of estimators N is 100. The number of max features per node M is 50. The minimum number of samples required to split is 50.

[4] The F1 score is the harmonic average of precision and recall and is often used as a single measurement that conveys the balance between precision and recall. It is widely used in evaluating the performance of algorithms (see, for example, Baeza-Yates and Ribeiro-Neto 1999).

[5] The statistical yearbook of year A records the latest available data before year A.

[6] We thank the associate editor and the review team for suggesting a direct analysis of the performance of doctors.

[7] We are grateful to an anonymous referee for pointing us to the science literature on human–algorithm collaboration complementarity.

References

Adepoju I-OO, Albersen BJA, De Brouwere V, van Roosmalen J, Zweekhorst M (2017) mHealth for clinical decision-making in sub-Saharan Africa: A scoping review. *JMIR mHealth uHealth* 5(3):e38.

Alur R, Raghavan M, Shah D (2024) Human expertise in algorithmic prediction. Globerson A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak J, Zhang C, eds. *Adv. Neural Inform. Processing Systems*, vol. 37 (Curran Associates, Inc., Red Hook, NY), 138088–138129.

Alur R, Laine L, Li DK, Raghavan M, Shah D, Shung D (2023) Auditing for human expertise. Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, eds. *Adv. Neural Inform. Processing Systems*, vol. 36 (Curran Associates, Inc., Red Hook, NY), 79439–79468.

Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, Tse D, et al. (2019) End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine* 25(6):954–961.

Ashar YK, Andrews-Hanna JR, Dimidjian S, Wager TD (2017) Empathic care and distress: Predictive brain markers and dissociable brain systems. *Neuron* 94(6):1263–1273.

Athey S, Imbens GW (2019) Machine learning methods that economists should know about. *Annual Rev. Econom.* 11:685–725.

Athey S, Wager S (2021) Policy learning with observational data. *Econometrica* 89(1):133–161.

Baeza-Yates R, Ribeiro-Neto B (1999) *Modern Information Retrieval*, vol. 463 (ACM Press, New York).

Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld D (2021) Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proc. 2021 CHI Conf. Human Factors Comput. Systems*, vol. 81 (Association for Computing Machinery, New York), 1–16.

Berg T, Burg V, Gombović A, Puri M (2020) On the rise of FinTechs: Credit scoring using digital footprints. *Rev. Financial Stud.* 33(7):2845–2897.

Berk R (2017) An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *J. Experiment. Criminology* 13(2):193–216.

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2021) Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* 50(1):3–44.

Breiman L (2001) Random forests. *Machine Learn.* 45(1):5–32.

Brennan M, Puri S, Ozrazgat-Baslanti T, Feng Z, Ruppert M, Hashemighouchani H, Momcilovic P, Li X, Wang DZ, Bihorac A (2019) Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: A pilot usability study. *Surgery* 165(5):1035–1045.

Bulathwela S, Pérez-Ortiz M, Holloway C, Shawe-Taylor J (2021) Could AI democratise education? Socio-technical imaginaries of an edtech revolution. Preprint, submitted December 3, https://arxiv.org/abs/2112.02034.

Chalfin A, Danieli O, Hillis A, Jelveh Z, Luca M, Ludwig J, Mullainathan S (2016) Productivity and selection of human capital with machine learning. *Amer. Econom. Rev.* 106(5):124–127.

Chan DC, Gentzkow M, Yu C (2022) Selection with variation in diagnostic skill: Evidence from radiologists. *Quart. J. Econom.* 137(2):729–783.

Chan S, Reddy V, Myers B, Thibodeaux Q, Brownstone N, Liao W (2020) Machine learning in dermatology: Current applications, opportunities, and limitations. *Dermatology Therapy* 10(3): 365–386.

Currie J, MacLeod WB (2017) Diagnosing expertise: Human capital, decision making, and performance among physicians. *J. Labor Econom.* 35(1):1–43.

Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, Ko J, et al. (2022) Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci. Adv.* 8(31):eabq6147.

De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, Askham H, et al. (2018) Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* 24(9):1342–1350.

Donahue K, Chouldechova A, Kenthapadi K (2022) Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. *Proc. 2022 ACM Conf. Fairness Accountability Transparency* (Association for Computing Machinery, New York), 1639–1656.

Dwyer DB, Falkai P, Koutsouleris N (2018) Machine learning approaches for clinical psychology and psychiatry. *Annual Rev. Clinical Psych.* 14:91–118.

Elliott G, Lieli RP (2013) Predicting binary outcomes. *J. Econometrics* 174(1):15–26.

Erickson BJ, Korfiatis P, Akkus Z, Kline TL (2017) Machine learning for medical imaging. *Radiographics* 37(2):505–515.

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

Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, New York).

Fekadu R, Getachew A, Tadele Y, Ali N, Goytom I (2022) Machine learning models evaluation and feature importance analysis on NPL dataset. Preprint, submitted August 28, https://arxiv.org/abs/2209.09638.

Feng K, Hong H (2024) Statistical inference of optimal allocations I: Regularities and their implications. Preprint, submitted March 27, https://arxiv.org/abs/2403.18248.

Feng K, Hong H, Tang K, Wang J (2022) Properties of ROC curves. Preprint, submitted May 30, https://dx.doi.org/10.2139/ssrn.3382962.

Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A (2022) Predictably unequal? The effects of machine learning on credit markets. *J. Finance* 77(1):5–47.

Fuster A, Plosser M, Schnabl P, Vickery J (2019) The role of technology in mortgage lending. *Rev. Financial Stud.* 32(5):1854–1899.

Gillis T, McLaughlin B, Spiess J (2021) On the fairness of machine-assisted human decisions. Preprint, submitted October 28, https://arxiv.org/abs/2110.15310.

Guo J, Li B (2018) The application of medical artificial intelligence technology in rural areas of developing countries. *Health Equity* 2(1):174–181.

Han SS, Park I, Eun Chang S, Lim W, Kim MS, Park GH, Chae JB, Huh CH, Na J-I (2020) Augmented intelligence dermatology: Deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J. Investigative Dermatology* 140(9):1753–1761.

Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY (2019) Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine* 25(1):65–69.

Huang S-C, Pareek A, Seyyedi S, Banerjee I, Lungren MP (2020) Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digital Medicine* 3(1):136.

Iakovlev A, Liang A (2024) The value of context: Human versus black box evaluators. Preprint, submitted February 17, https://arxiv.org/abs/2402.11157.

Jin W, Fatehi M, Guo R, Hamarneh G (2024) Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. *Artificial Intelligence Medicine* 148:102751.

Johnson EM, Rehavi MM (2016) Physicians treating physicians: Information and incentives in childbirth. *Amer. Econom. J. Econom. Policy* 8(1):115–141.

Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E, Dudley JT (2018) Artificial intelligence in cardiology. *J. Amer. College Cardiology* 71(23):2668–2679.

Kahneman D, Klein G (2009) Conditions for intuitive expertise: A failure to disagree. *Amer. Psych.* 64(6):515–526.

Kawaguchi K (2021) When will workers follow an algorithm? A field experiment with a retail business. *Management Sci.* 67(3): 1670–1695.

Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, et al. (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172(5):1122–1131.

Kitagawa T, Tetenov A (2018) Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86(2):591–616.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. *Quart. J. Econom.* 133(1):237–293.

Kotz S, Balakrishnan N, Johnson NL (2019) *Continuous Multivariate Distributions, Volume 1: Models and Applications*, vol. 334 (John Wiley & Sons, Hoboken, NJ).

Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, Cai W, et al. (2019) Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Medicine* 25(3):433–438.

Long E, Lin H, Liu Z, Wu X, Wang L, Jiang J, An Y, et al. (2017) An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nature Biomedical Engrg.* 1(2):0024.

Lopez C, Sautmann A, Schaner S (2018) The contribution of patients and providers to the overuse of prescription drugs. NBER Working Paper No. w25284, National Bureau of Economic Research, Cambridge, MA.

Manski CF (2018) Credible ecological inference for medical decisions with personalized risk assessment. *Quant. Econom.* 9(2):541–569.

Mbakop E, Tabord-Meehan M (2021) Model selection for treatment choice: Penalized welfare maximization. *Econometrica* 89(2): 825–848.

McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, et al. (2020) International evaluation of an AI system for breast cancer screening. *Nature* 577(7788):89–94.

Mondal H, Mondal S, Singla RK (2023) Artificial intelligence in rural health in developing countries. Chatterjee JM, Saxena SK, eds. *Artificial Intelligence in Medical Virology* (Springer Nature, Singapore), 37–48.

Mullainathan S, Obermeyer Z (2022) Diagnosing physician error: A machine learning approach to low-value health care. *Quart. J. Econom.* 137(2):679–727.

Mullainathan S, Spiess J (2017) Machine learning: An applied econometric approach. *J. Econom. Perspect.* 31(2):87–106.

O'Neil C (2017) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, New York).

Peng S, Liu Y, Lv W, Liu L, Zhou Q, Yang H, Ren J, et al. (2021) Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: A multicentre diagnostic study. *Lancet Digital Health* 3(4):e250–e259.

Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY (2017) Cardiologist-level arrhythmia detection with convolutional neural networks. Preprint, submitted July 6, https://arxiv.org/abs/1707.01836.

Rambachan A, Roth J (2019) Bias in, bias out? Evaluating the folk wisdom. Preprint, submitted September 18, https://arxiv.org/abs/1909.08518.

Rambachan A, Kleinberg J, Ludwig J, Mullainathan S (2020) An economic perspective on algorithmic fairness. *AEA Papers Proc.*, vol. 110, 91–95.

Ren H, Wang J, Zhao WX, Wu N (2021) RAPT: Pre-training of time-aware transformer for learning robust healthcare representation. *Proc. 27th ACM SIGKDD Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 3503–3511.

Sadhwani A, Giesecke K, Sirignano J (2021) Deep learning for mortgage risk. *J. Financial Econometrics* 19(2):313–368.

Sajjadiani S, Sojourner AJ, Kammeyer-Mueller JD, Mykerezi E (2019) Using machine learning to translate applicant work history into predictors of performance and turnover. *J. Appl. Psych.* 104(10):1207–1225.

Studdert DM, Mello MM, Sage WM, DesRoches CM, Peugh J, Zapert K, Brennan TA (2005) Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. *JAMA* 293(21):2609–2617.

Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J (2023) From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell* 186(8):1772–1791.

Tian F, Liu D, Wei N, Fu Q, Sun L, Liu W, Sui X, et al. (2024) Prediction of tumor origin in cancers of unknown primary origin with cytology-based deep learning. *Nature Medicine* 30(5):1309–1319.

Trivedi A, Mukherjee S, Tse E, Ewing A, Ferres JL (2019) Risks of using non-verified open data: A case study on using machine learning techniques for predicting pregnancy outcomes in India. Preprint, submitted October 4, https://arxiv.org/abs/1910.02136.

Uhm K-H, Jung S-W, Choi MH, Shin H-K, Yoo J-I, Oh SW, Kim JY, et al. (2021) Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography. *NPJ Precision Oncology* 5(1):54.

Vaccaro M, Almaatouq A, Malone T (2024) When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behav.* 8(12):2293–2303.

Vallee B, Zeng Y (2019) Marketplace lending: A new banking paradigm? *Rev. Financial Stud.* 32(5):1939–1982.

Wang Z, Wei L, Xue L (2024) Overcoming medical overuse with AI assistance: An experimental investigation. Preprint, submitted May 17, https://arxiv.org/abs/2405.10539.

Wang J, Gallo E, Zhang W, Tang K, Hong H (2023) Diagnosing with the help of artificial intelligence. Working paper, Beihang University, Beijing.

Yang J, Xie M, Hu C, Alwalid O, Xu Y, Liu J, Jin T, et al. (2021) Deep learning for detecting cerebral aneurysms with CT angiography. *Radiology* 298(1):155–163.

Yoon J, Kang C, Kim S, Han J (2022) D-vlog: Multimodal vlog dataset for depression detection. *Proc. AAAI Conf. Artificial Intelligence*, vol. 36 (PKP Publishing Services Network, Vancouver), 12226–12234.

Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P (2024) Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine* 30(3):837–849.

Zeng J, Ustun B, Rudin C (2017) Interpretable classification models for recidivism prediction. *J. Roy. Statist. Soc. Ser. A Statist. Soc.* 180(3):689–722.