Doctor Decision Making and Patient Outcomes*

Janet Currie[†], W. Bentley MacLeod[‡] and Kate Musen[§]

March 3, 2025

Abstract

Doctors often treat similar patients differently, affecting health and spending. We review the recent literature on physician decision making through the lens of a model that incorporates doctors diagnostic and procedural skills, beliefs, and incentives as well as differences in patient pools. The quality of decision making is affected by training, experience, peer effects, financial incentives, and time constraints. Interventions to improve decision making include providing information, guidelines, and technologies like electronic medical records and algorithms. Economists have made progress in understanding doctor decision making, but our ability to apply that knowledge to improve health care is still limited.

1 Introduction

Doctors facing similar patients often make different treatment choices, and these can have large consequences for health outcomes and health care spending. Badinski et al. (2023) show that roughly a third of regional differences in the healthcare utilization of elderly Americans is explained by differences in average physician treatment intensity. Health care accounts for almost 20% of U.S. GDP, and many observers believe that much of that spending is misdirected, wasted, or even harmful (Chandra and Skinner (2012), Cutler (2014)). A rapidly growing literature focuses on understanding the sources of this variation. We are all health care consumers, so the question of what drives doctor decision making is of intrinsic interest. However, understanding doctor decision-making could also shed light on the behavior of other experts such as lawyers, top managers, or even professors, who share characteristics such as intensive training, considerable autonomy, and a sometimes uncertain relationship between inputs and outputs.

This paper seeks to organize the recent literature (since 2010) on physician decision making by looking at it through the lens of a model that has several key elements. First, doctors care about patients, but they are influenced by their beliefs about appropriate care, time constraints, and profit motives, all of which can vary across doctors. Hence, doctors are imperfect agents from the point of view of patients, given that they care about other considerations in addition to patient utility. Second, doctors' skill levels vary. We distinguish between skill involved in deciding what to do (diagnosis) and procedural skill, defined as skilled execution of a given decision. Third, patients care about medical outcomes, and other factors including quality of life and out-of-pocket costs. Both doctors and patients may have strong beliefs about treatments (e.g., doctors

*We would like to thank David Chan, Jonathan Gruber, Amanda Kowalski, David Romer, Tim Wang and four anonymous referees for helpful comments. Kate Musen gratefully acknowledges support from the National Science Foundation (Grant Number DGE2036197).

[†]Princeton University and NBER

[‡]Yale University and NBER

[§]Columbia University

may have been trained to think that a procedure is necessary, and patients may believe, for example, that vaccines are harmful).[1] All of these factors mean that patients with identical conditions can end up being treated differently.

Table 1 describes a number of studies demonstrating that physicians often treat similar patients so differently that they can be said to have distinct "practice styles". For example, Berndt et al. (2015) study concentration in the way that doctors prescribe anti-psychotics and show that typically, two-thirds of a doctor's prescriptions are for the same drug, and that crucially, doctors have different favorite drugs. Cutler et al. (2019) use Medicare claims data to identify "cowboys," who recommend aggressive treatments that go beyond clinical guidelines and "comforters," who recommend palliative care for severely ill patients. Focusing on elderly heart attack (i.e. acute myocardial infarction or AMI) patients, they find that a one standard deviation increase in the share of doctors who are cowboys leads to a 13% increase in annual spending, whereas a one standard deviation increase in the share of comforters leads to a small decrease in annual spending. Notably, neither share is associated with changes in survival probabilities.[2] Fadlon and van Parys (2020) look at patients who switch providers after their primary care physician retired or moved away. They find that changing to a provider who spends more on primary care increases spending on primary care, which they interpret as evidence of distinct practice styles. Ahammer and Schober (2020) show similar results in the Austrian context. Marquardt (2021) examines variation in diagnoses of ADHD and finds that a one standard deviation increase in physician "intensity" (measured as the intercept in a doctor-specific regression) increases the probability a patient is diagnosed by 22.45.

The model outlined in the next Section builds on work in three of the papers shown in Table 1— Abaluck et al. (2016), Currie and MacLeod (2017), and Chan et al. (2022) to provide a framework to think about alternative reasons for the observed variation in physician decision making and about interventions that have been suggested in an effort to improve outcomes. The literature on health disparities discussed in Section 3, shows that an individual physician may vary treatment based on characteristics of the patient that are unrelated to their health status, illustrating the role that idiosyncratic physician preferences play in treatment decisions. Economic considerations that affect the quality of decision making include financial incentives, experience, training, peer effects, and time constraints, are discussed in Section 4. Another branch of the literature asks whether decision making can be improved through informational interventions, guidelines, or the use of technology including algorithmic decision tools. These are discussed in Section 5 of the paper.

Understandably, most of the studies we review focus on the role of a single explanatory factor, although this often requires strong assumptions about the other factors. Our first objective is to make these assumptions more explicit. Second, we try to connect aspects of the decision process that are typically studied in isolation, such as the relationship between doctor skill and thresholds for choosing aggressive procedures. Third, we offer an empirical assessment of what we have learned to date about doctor decision making and suggestions for further research.

[1] One of the most famous examples of the persistence of erroneous beliefs about the efficacy of treatment has to do with blood letting, a treatment that persisted for centuries even though it is now known to be more likely to harm than help patients. See Parapia (2008) for an interesting history of attitudes toward blood-letting as a medical practice. In an era when many sick patients died, a few patients surviving after blood letting might have reinforced doctor beliefs in the benefits of the treatment.

[2] Clemens et al. (2024) look at the same doctors as Cutler et al. (2019) and find that doctor preferences have less impact on practice patterns in the privately insured population than they do in Medicare. They hypothesize that this difference reflects greater variation in prices across private insurance plans, since prices also influence doctor behavior.

2 A Simple Model of Physician Behavior and Patient Outcomes

This Section sketches a simple model of physician decision making. The technical details and proofs are relegated to the Appendix. Consider patient $i \in \mathcal{N}_j$ who seeks treatment from physician j where $\mathcal{N}_j$ denotes the set of patients seen by j. In what follows, any variable that changes with the patient is subscripted with i, and similarly, variables that vary by physician are subscripted by j.

Physician j can treat patient i with one of two treatments, a nonintensive treatment ($t_{ij} = NI$) or an intensive treatment ($t_{ij} = I$). For example, Chandra and Staiger (2007) and Currie et al. (2016) consider heart patients where the choice is cardiac catheterization (the intensive procedure) versus medical (i.e., drug) management. Currie and MacLeod (2017) study childbirth, where a vaginal delivery is the nonintensive procedure and a C-section is the invasive procedure. In Abaluck et al. (2016), the "intensive" (or at least more expensive) procedure is to test a patient for a pulmonary embolism and the nonintensive alternative is not to test.

Assume that there is a best medical choice for patient i, given by their *unobserved* state $\alpha_i \in \{L, H\}$. If $\alpha_i = L$, then the patient is low risk, and the nonintensive treatment is preferred, while $\alpha_i = H$ implies that the patient is high risk, and the intensive treatment is more appropriate.

The economics literature on doctor decision making, such as Chandra and Staiger (2007), assumes that doctors make the choice that maximizes their own expected utility. The utility for doctor j giving a patient of type α treatment t is:

$$U_{\alpha tj} = u_{\alpha tj} + \delta_{tj}, \tag{1}$$

where $u_{\alpha tj}$ is the expected medical benefit to a patient of type $\alpha \in \{L, H\}$ getting treatment $t \in \{NI, I\}$ from doctor j. The medical benefit to the patient, $u_{\alpha tj}$, can differ by doctor, depending on the doctor's *procedural skill*. For example, if a doctor is a skilled surgeon, then the result of a difficult surgery may well be much better than if the same procedure was performed by a mediocre surgeon. Additional factors that affect treatment, such as doctor payments for administering the treatment and variations in the cost of treatment, are captured by δ_{tj}. The δ_{tj} are normalized so that $\delta_{NIj} = 0$, while δ_{Ij} represents the pecuniary returns to intensive treatment. It is possible for $\delta_{Ij} < 0$ if, for example, hospital or insurance plans set rewards so as to discourage use of the intensive procedure.

If the patient is low risk, then the nonintensive treatment will have a higher medical benefit ($u_{LNIj} > u_{LIj}$), while for type $\alpha = H$ the intensive treatment is more medically beneficial ($u_{HIj} > u_{HNIj}$). Let the increase in doctor utility for patients getting the appropriate treatment be:

$$\Delta_{HIj} = \{U_{HIj} - U_{HNIj}\} = u_{HIj} - u_{HNIj} + \delta_{Ij},$$

$$\Delta_{LNIj} = \{U_{LNIj} - U_{LIj}\} = u_{LNIj} - u_{LIj} - \delta_{Ij}.$$

Doctors have ex ante beliefs regarding the appropriate treatment for patients in their pool of potential patients:

$$p_{Hj} = \Pr\left[\alpha = H | j\right],$$

while the ex ante probability estimate that $\alpha_i = L$ is $p_{Lj} = 1 - p_{Hj}$.[3]

[3] It is not necessarily the case that doctors know the true distribution of types, hence one cannot assume $p_{Hj} = \Pr\left[\alpha = H | i \in \mathcal{N}_j\right]$.

The patient's true condition is α_i, but doctor j's diagnosis is based on a noisy signal that is correlated with patient i's condition (whether α_i is H or L):

$$T_{ij} = \begin{cases} 1 + \epsilon/\gamma_j, & if\,\alpha_i = H, \\ -1 + \epsilon/\gamma_j & if\,\alpha_i = L. \end{cases} \tag{2}$$

where $\epsilon \sim N(0,1)$ and γ_j is diagnostic skill.[4] The mean of the signal is 1 when $\alpha_i = H$, and -1 when $\alpha_i = L$. An increase in diagnostic skill reduces the variance of the signal, reducing the probability of a misdiagnosis. Although diagnostic skill is often ignored by economists, the National Academy of Sciences[5] notes that diagnostic errors—which they define as inaccurate or delayed diagnoses—are frequent, affecting 5% of American outpatients annually, contributing to between 6% and 17% of hospital adverse events, and ultimately leading to 10% of patient deaths. Diagnostic errors are also a leading cause of successful medical malpractice cases.

The signal T_{ij} is increasing in α_i so, as in Chandra and Staiger (2007), it follows that the decision rule for the treatment $t_{ij} \in \{NI, I\}$ takes the form:[6]

$$t_{ij} = \begin{cases} I, & T_{ij} \geq \tau_j, \\ NI, & T_{ij} < \tau_j, \end{cases}$$

where τ_j is the doctor's *decision threshold* for deciding when to implement the intensive treatment. Increasing the threshold reduces the probability that the intensive treatment is chosen. Chandra and Staiger (2007) assume that in areas where doctors do a lot of the intensive procedure, they become more skilled at the intensive procedure and less skilled at the nonintensive procedure, which causes the threshold for the intensive procedure to fall, leading to more intensive procedures. This Section extends their model by adding diagnostic skill.

The quality of diagnosis is measured by the likelihood that a patient is assigned to the correct medical treatment. There are two measures of performance that correspond to whether patients correctly or incorrectly receive the intensive treatment. The first is the probability that a patient i of type $\alpha_i = H$ receives the appropriate treatment. The second measure is the probability that patient i of type $\alpha_i = L$ receives the inappropriate intensive treatment. Since there is uncertainty in the doctor's mind regarding the true state, increasing the probability of the type H patients getting the intensive treatment will mechanically have the negative consequence of increasing the probability that patients of type L get the inappropriate intensive treatment.

This trade-off is illustrated in Figure 1 which shows a plot of the probability of appropriate versus inappropriate intensive treatment for different levels of diagnostic skill, γ_j. This curve is the well-known receiver-operator curve, or ROC from machine learning, where the probability of appropriate intensive treatment for a high need patient is the *True Positive Rate* or $TPR_j = \Pr[t_{ij} = I_i | \alpha_i = H, j]$ while the probability of inappropriate intensive treatment for a low need patient is the *False Positive Rate* or $FPR_j = \Pr[t_{ij} = I_i | \alpha_i = L, j]$.[7] Chan et al. (2022) observe that when the ROC curve of one decision maker is above another, they are processing information more efficiently (see Remark I in Section II.B).

[4]The assumption that ϵ is Normal allows for elegant closed form solutions and provides intuition that holds for many cases considered in the literature that do not assume a Normal distribution.

[5]Balogh et al. (2015).

[6]See Section A in Chandra and Staiger (2007) and Abaluck et al. (2016).

[7]See Fawcett (2006).

As γ_j increases, the frontier moves up and left. The top left corner represents perfect diagnosis—the patient receives the intensive treatment if and only if they are of type $\alpha_i = H$. Conversely, as γ_j approaches zero, the frontier approaches the dashed 45 degree line. The decision threshold τ_j defines a point on the diagnostic frontier. As τ_j increases, the doctor has a higher threshold for performing the intensive procedure, so the probability of intensive treatment falls for all patients.

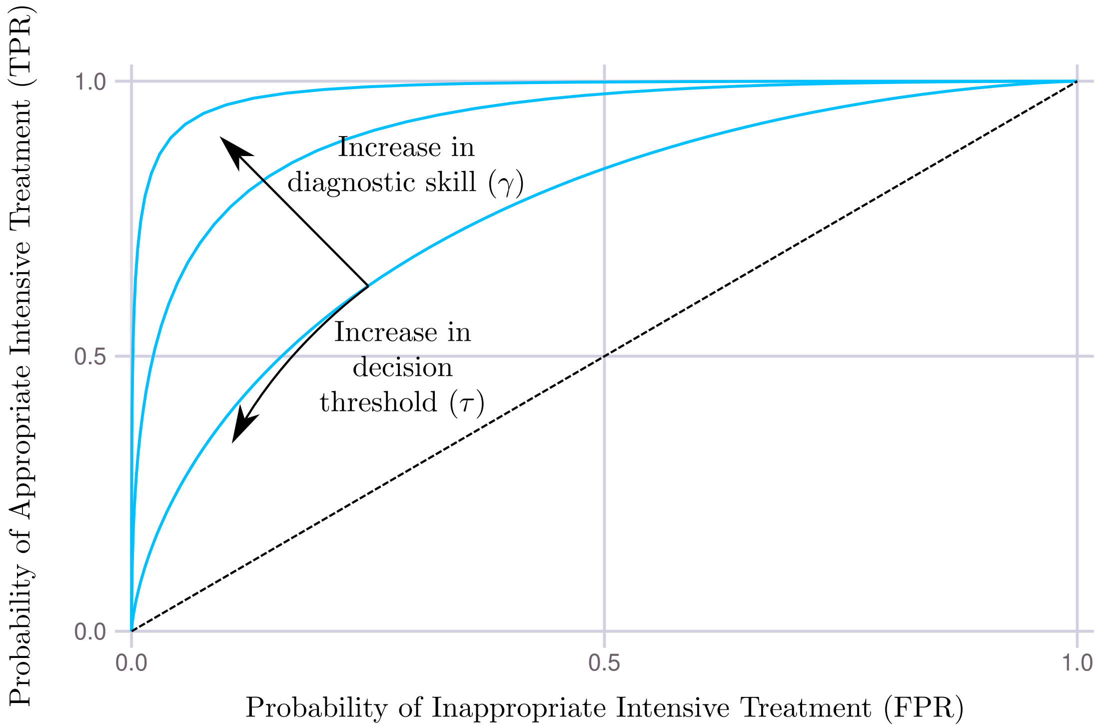

Figure 1: Effect of Diagnostic Skill

Given this set up, the doctor's utility maximizing threshold τ_j^* is:

$$\tau_j^* = b_j^* / \gamma_j^2, \tag{3}$$

where $b_j^* \equiv \left(\ln\left(\Delta_{LNIj}/\Delta_{HIj}\right) + \ln\left(p_{Lj}/p_{Hj}\right)\right)/\,2$ is the unadjusted decision *threshold* that summarizes physician preferences, while τ_j^* is the utility maximizing decision threshold taking into account diagnostic skill.[8]

Equation (3) shows that the decision threshold depends on diagnostic skill, γ_j, the relative effectiveness of nonintensive and intensive treatments for the two types of patients, $\Delta_{LNIj}/\Delta_{HIj}$, and the doctor's beliefs about the relative proportion of patient types, p_{Lj}/p_{Hj}, in their patient pools. If a doctor believes that most patients need nonintensive treatment, the doctor will adopt a higher decision threshold for the use of intensive treatment compared to a doctor who believes the reverse. If the relative benefit from intensive treatment is higher, doctors will adopt a *lower* decision threshold resulting in more use of the intensive procedure. If the pecuniary benefit δ_{Ij} for selecting the intensive treatment is sufficiently small then $\Delta_{HIJ} < 0$, and doctor j chooses only the nonintensive procedure for all patients. Conversely, if the pecuniary gain to the intensive

[8]See Propositions 1 and 2 in the Appendix.

treatment is sufficiently large that $\Delta_{LNIj} < 0$, then the intensive treatment is selected regardless of the signal.

When neither of these cases hold, then greater diagnostic skill, γ_j makes the doctor's beliefs about the distribution of patient types and the expected relative benefits of the procedures less important. This is because a doctor with perfect diagnostic skill observes the patient's true condition, and chooses the procedure appropriate for that patient. When diagnostic skill falls, physicians choose the treatment that they believe is ex ante medically appropriate for most patients, and more patients receive the same treatment. Note that while this behavior would increase the within-doctor uniformity of treatment, it could increase the variance in across doctor-behavior, depending on the distribution of doctor beliefs.[9]

These results are illustrated in Figure (2). It illustrates outcomes for two doctor types with different practice styles:

- A cautious doctor (C), or "comforter" in the Cutler et al. (2019) terminology, is one who is more likely to give a nonintensive treatment. In this case, the shift parameter is $b_C = \log\left(\frac{\Delta_{0NIC}}{\Delta_{1IC}} \times \frac{p_{0C}}{p_{1C}}\right) > 0$. The decision threshold is at the point where the slope, which in this case is greater than one $\left(\frac{\Delta_{0NIC}}{\Delta_{1IC}} \times \frac{p_{0C}}{p_{1C}} > 1\right)$, is tangent to the diagnostic frontier. The points τ^*_{CH}, τ^*_{CM} and τ^*_{CL}, correspond to cautious doctors with high, medium, and low diagnostic skills, respectively.

- An aggressive doctor (A), or "cowboy" in the Cutler et al. (2019) terminology, is one who is more likely to do the intensive treatment. In this case the shift parameter is $b_A = \log\left(\frac{\Delta_{0NIA}}{\Delta_{1IA}} \times \frac{p_{0A}}{p_{1A}}\right) < 0$. The decision threshold is at the point where the slope, which in this case is less than one $\left(\frac{\Delta_{0NIC}}{\Delta_{1IC}} \times \frac{p_{0C}}{p_{1C}} < 1\right)$, is tangent to the diagnostic frontier. The points τ^*_{AH}, τ^*_{AM} and τ^*_{AL} correspond to doctors with high, medium, and low diagnostic skill, respectively.

The figure shows that even if doctors base their decisions on what is medically appropriate for the patient, it is still the case that ex ante beliefs about the probability that the nonintensive treatment is appropriate (p_{Lj}/p_{Hj}) affect their choices.

This stylized model builds on the framework developed in the machine learning literature.[10] It illustrates how doctor decisions depend on a number of factors that may or may not be observed. These factors include the characteristics of the population seeking treatment, the doctor's beliefs regarding this population, their ability to correctly update these beliefs given the available information, the costs and benefits from treatment for both types of patients (which will depend in part on the doctor's procedural skill), and the pecuniary rewards that the doctor gets for making a particular choice.

Outcomes for both types of patients can improve with an increase in diagnostic skill. Higher γ_j, always results in an increase in TPR-FPR. This quantity is the difference between the probability that high-risk patients will get the high-intensity treatment, and the probability that low-risk patients will incorrectly get the high-intensity treatment. For clarity, we have assumed normally distributed errors so that greater diagnostic skill always results in improvements for both types of patients. Chan et al. (2022) and Rambachan (2024) explore more general models of diagnostic skill and provide conditions under which these results generalize.

[9] See Proposition 3 in the Appendix.

[10] See Feng et al. (2023) for an explicit application of machine learning to doctor decision making, including a discussion of how to estimate ROC curves.

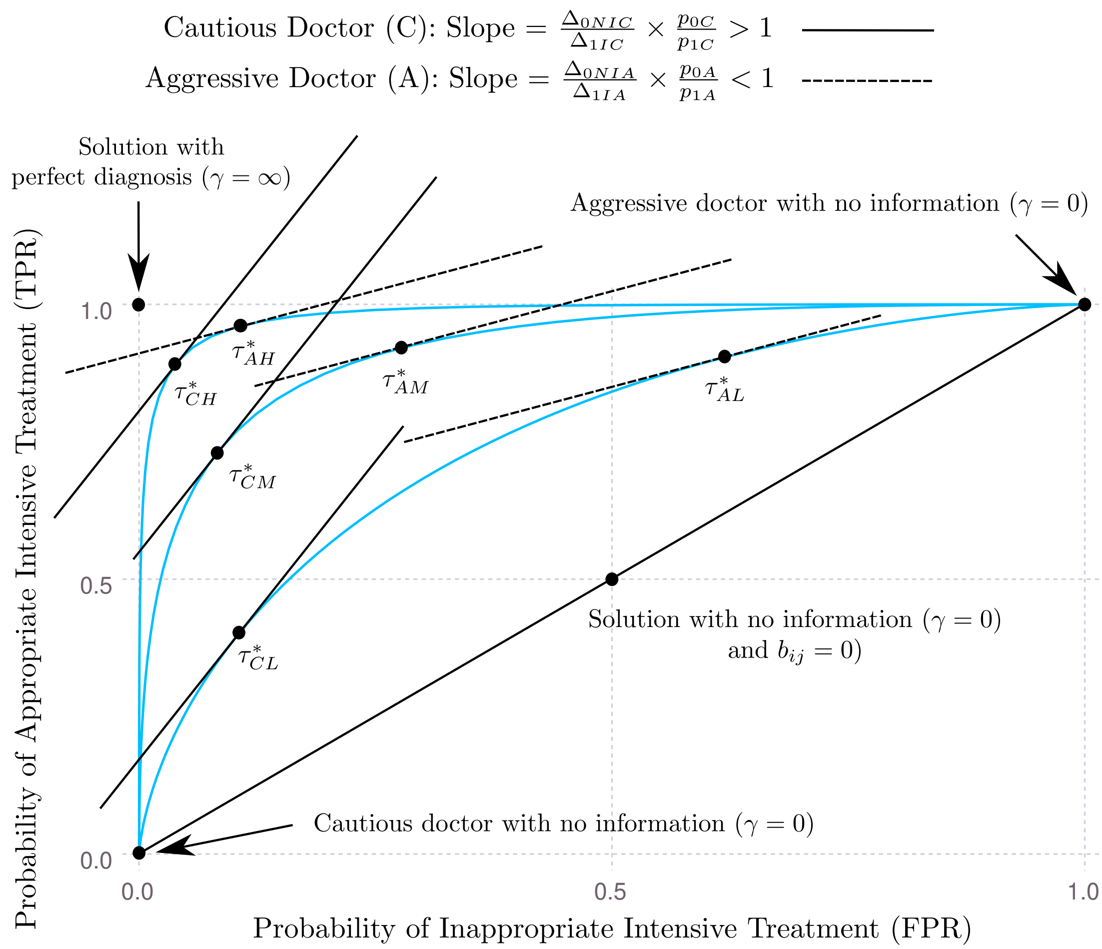

Figure 2: Doctor's Diagnostic Rule

2.1 Identifying doctor diagnostic thresholds, diagnostic skill, and procedural skill from data

Studies of doctor decision making typically have data that include information about the primary treatment choice ($t_{ij} \in \{NI, I\}$), some measures of patient health following treatment, and some information about patient type from medical records. In this section, we discuss three papers that illustrate some of the challenges one faces when estimating the quality of physician decision making using such data.

We begin with a discussion of Chan et al. (2022), who exploit the random assignment of patients who may have pneumonia to radiologists. The radiologist's problem is to diagnose whether the patient has pneumonia or not, where patients with pneumonia will be admitted to hospital and those without will be sent home. Even though checking x-rays for signs of pneumonia is a very routine task for radiologists, there appears to be significant variation in their diagnostic skill. This result suggests that we might expect to find even more variation in diagnostic skill in other, arguably more complex, medical contexts. However, in most settings, people have some choice over providers, so that one cannot assume that patients are randomly assigned to doctors. Hence, we next discuss Abaluck et al. (2016) and Currie and MacLeod (2017) , which illustrate two approaches to identifying skill from field data in the absence of random assignment.

Chan et al. (2022) build on a literature that exploits the random assignment of individuals to judges to estimate biases in judicial decision making.[11] See also Rambachan (2024) for a recent extension of these identification results. Arnold et al. (2022) look at a judge's decision to grant bail or not. Bail is not granted if the judge believes there is a high probability that the individual will re-offend. The challenge is that when bail is not granted then one does not know whether the person would have re-offended. Arnold et al. (2022) introduce a hierarchal marginal treatment effect model that allows them to identify judge decision skill, in addition to the decision threshold.

A unique feature of Chan et al. (2022)'s data is that patients with missed pneumonia diagnoses are likely to return to hospital, which allows them to measure the fraction of cases that each radiologist missed. In contrast, it is more difficult to say whether a patient who received a C-section would have been better off without one since one does not see the counterfactual for each patient.[12]

Chan et al. (2022) show that this information is sufficient to identify each doctor's probability of recommending appropriate intensive treatment, or TRP_j and the probability of inappropriately recommending intensive treatment or FPR_j.[13] Given (FPR_j, TRP_j) for each physician, one can use the model to derive both diagnostic skill and the decision threshold from the following equation:

$$TPR\left(\tau_j, \gamma_j\right) \equiv \Pr\left[T_{ij} \geq \tau_j | \alpha_i = H\right] = F\left(\gamma_j\left(1 - \tau_j\right)\right), \tag{4}$$

where $F(\cdot)$ is the Normal cumulative probability distribution, and

$$FPR\left(\tau_j, \gamma_j\right) \equiv \Pr\left[T_{ij} \geq \tau_j | \alpha = L\right] = F\left(\gamma_j(-1 - \tau_j)\right). \tag{5}$$

Hence, given $TPR_j \in (0,1)$, $FPR_j \in (0,1)$, and $TPR_j > FPR_j$ there is a unique solution for $\tau_j \in (-\infty, \infty)$ and $\gamma_j > 0$ solving (4-5).[14]

Like Figures (1—2), Figure (3), taken from Chan et al. (2022), illustrates the relationship between appropriate and inappropriate testing. Each point corresponds to the average true positive and the false positive rate of a radiologist for the population of patients that they treat. If doctors only varied in terms of their decision thresholds, then all the points would lie on the same curve. Similarly, if all the doctors differed only in terms of diagnostic skill, then the points would follow a line such as that connecting the points τ^*_{AH}, τ^*_{AM} and τ^*_{AL} in Figure 2. Instead, these data suggest a great deal of variation in diagnostic skill as well as some variation in thresholds.

In addition to the random assignment of patients to doctors and the fact that they can observe ex post whether the doctor made a mistake, another valuable feature of Chan et al. (2022)'s setting is that in the case of a radiologist interpreting an x-ray image, it is reasonable to assume that variation in outcomes is due only to diagnostic skill. In many other medical settings there is a meaningful distinction between deciding when an intensive procedure is appropriate, and actually performing the intensive procedure. Thus, the Chan et al. (2022) setting excludes three factors that are likely to be important in other medical settings:

[11]See Chyn et al. (2024) for an extensive review of the literature using random assignment. They point out that even with randomization, there are situations in which estimates of the treatment effect are biased. They discuss some of the techniques used to address these issues.

[12]It seems that Abaluck et al. (2016) could have looked for missed diagnoses of pulmonary embolism (PE) in their data by looking for people who returned to hospital with complications of PE. However, a large number of people with PE die outside the hospital who are not in the authors' hospital claims data.

[13]The details are in Section C of the online appendix to Chan et al. (2022).

[14]Proposition 4 in appendix. See also Section E of the online appendix to Chan et al. (2022) for the derivation of a structural model building on this observation.

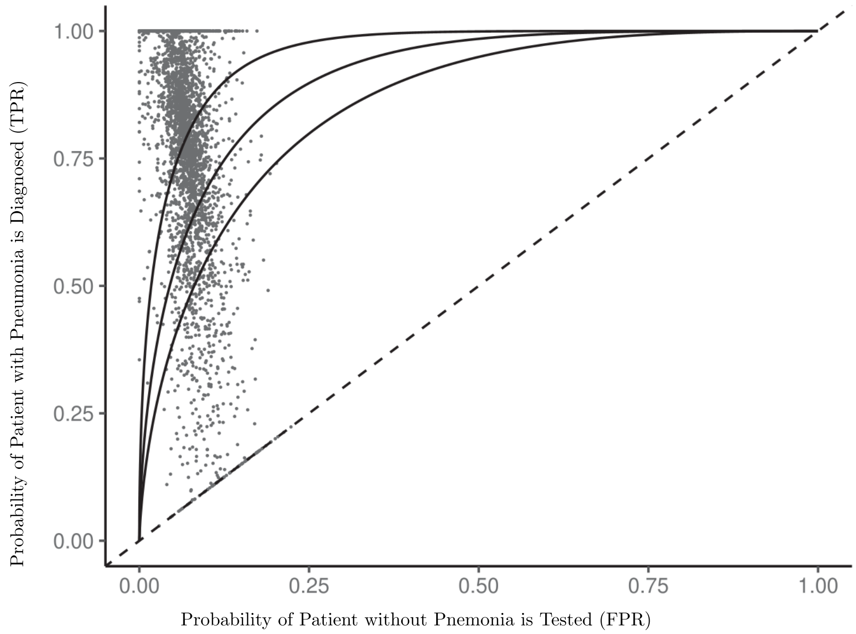

Figure 3: Distribution of Decision Thresholds and Diagnostic Skill for Radiologists (modified version of Figure V of Chan et al. (2022))
Note: Each point represents one radiologist.

selective matching of patients and doctors, the ability to observe ex post whether the doctor made an error, and the distinction between procedural and diagnostic skill.

Abaluck et al. (2016) is an example of a study that uses observational Medicare claims data to estimate doctors' decision thresholds. This widely used data source covers most U.S. elderly and hence provides a large, nationally representative sample of doctors and their patients. Abaluck et al. (2016) study doctors who order computerized tomography (CT) scans for patients suspected of having a life-threatening pulmonary embolism (PE). A near-definitive diagnosis can be made with a CT scan, but scans are expensive and expose patients to potentially harmful radiation, so it is possible to order too many scans.[15]

The lack of random assignment of patients to doctors is solved by making parametric assumptions regarding the likelihood that doctor j's patients have a PE. Specifically, it is assumed that the doctor's signal of patient condition is given by their estimate of patient i's probability of having a PE:

$$\begin{aligned} T_{ij} &= \Pr\left[\alpha = H | i, j\right] && (6) \\ &= \vec{x}_i \beta + a_j + \eta_{ij}, && (7) \\ &\equiv \rho_j\left(\vec{x}_i\right) + \eta_{ij}, && (8) \end{aligned}$$

where $\vec{x}_i$ is a vector of observed patient characteristics, and a_j is a doctor fixed effect. The doctor fixed effect, a_j, is the mean rate of PE for the patient population faced by doctor j. The error term, η_{ij}, reflects unobserved patient characteristics net of the average differences in the patient populations, and it is assumed to have a fixed distribution that can be estimated from the data.

The doctor orders a CT scan whenever $T_{ij} \geq \tau_j^*$, that is when they believe that the probability of a PE

[15]The authors do note that the downstream cancer risk from radiation exposure is less of a concern in the elderly population they study.

is greater than τ_j^*. This problem can be formulated as a standard selection model that can be estimated from the data:

$$\begin{aligned} T_{ij} - \tau_j^* &= \vec{x}_i \beta + a_j - \tau_j^* + \eta_{ij}, \\ &= \vec{x}_i \beta + \hat{a}_j + \eta_{ij}, \\ &\geq 0, \end{aligned}$$

where the distribution of η_{ij} is given by the cumulative distribution function $H(\cdot)$. This specification allows one to estimate an equation of the form:

$$\Pr[t_{ij} = I | \vec{x}_i, j] = 1 - H(\vec{x}_i \beta + \hat{a}_j). \tag{9}$$

Since both a_j and τ_j^* enter linearly, they cannot be separately identified. Abaluck et al. (2016) provide a clever solution to this problem. Given (9), they show that there is a selection function $\lambda(\cdot)$ such that:

$$\Pr[\alpha = H | \vec{x}_i, t_{ij} = I] = \tau_j^* + \lambda(\vec{x}_i \beta + \hat{a}_j). \tag{10}$$

Since the patients are tested if and only if the probability of a positive test is at least τ_j^*, the left-hand side of (10) is greater than or equal to τ_j^*; hence $\lambda(\cdot) \geq 0$. If a doctor has a sufficiently large number of patients, then many tested individuals will be on the threshold between being tested or not:

$$M_j = \{i \in \mathcal{N}_j | \lambda(\vec{x}_i \beta + \hat{a}_j) \approx 0, t_{ij} = I\}.$$

For these individuals the probability that they have a PE is exactly the decision threshold:

$$\tau_j^* = E\{\alpha_i = H | i \in M_j\}.$$

Abaluck et al. (2016) observe that a shortcoming of this approach is that the number of marginal patients may be small, which can result in an imprecisely measured decision threshold.

Having estimated the mean population risk, a_j, and decision threshold, τ_j^* for doctor j, Abaluck et al. (2016) then ask if the weights, β, used to estimate risk are correct. They do this by estimating a model for whether the patient has a PE including the selection term derived above and asking if the observables have additional explanatory power.[16] They find that, on average, doctors are using the wrong weights when deciding whether to order a test. They assume that all doctors use the same weights, that is, they all have similar diagnostic skills. Hence, by construction, variation in doctor behavior in their model comes from differences in doctor thresholds and patient pools.

The Appendix shows that one can compute the true positive rate, $TPR(\vec{x}_i, a_j, \tau_j)$, and false positive rate, $FPR(\vec{x}_i, a_j, \tau_j)$, given the Abaluck et al. (2016) model. In addition to illustrating that their model maps to a standard ROC curve, the Appendix illustrates that holding the patient pool fixed, doctors vary only in terms of their decision thresholds in their model. As both Chan et al. (2022) and Feng et al. (2023) observe, if in reality doctors can vary with regard to both the decision threshold and diagnostic skill, then different doctors generate different ROC curves.

Currie and MacLeod (2017) examine doctor thresholds for intensive procedures, diagnostic skill, and

[16] See equation (8) in Abaluck et al. (2016).

procedural skill using a dataset consisting of all births in New Jersey from 1997 to 2006 and focusing on C-section deliveries as the intensive procedure. To address the fact that women usually choose their ob-gyn practice, the authors use an instrumental variables strategy based on the fact that the majority of women choose a practice within a well-defined market. They then exploit the fact that there is variation in mean diagnostic skill, decision thresholds, and procedural skills across markets.

The doctor's decision is between a vaginal delivery (the nonintensive treatment) and cesarean section (CS: the intensive treatment). The doctor deciding on a CS will normally also perform it, but there is still a meaningful distinction between correctly diagnosing that someone needs a C-section and performing it well. Procedural skill will be reflected in the relative returns from treatment, $\Delta_{LNIj}/\Delta_{HIj}$. Doctors who are better at performing vaginal deliveries will have a higher Δ_{0NI}, while better surgeons have a higher Δ_{HIj}.

As in Abaluck et al. (2016) and Chan et al. (2022), one can use the vector of observed patient characteristics, $\vec{x}_i$ to estimate the patient's appropriateness for the intensive procedure and treat this estimated probability as an index of appropriateness, that is the medical benefit of the procedure. Consistent with the discussion above, we can let this index, $\rho(\vec{x}_i)$ be defined as the expected probability that patient i obtains a C-section in the population being studied. Over the period Currie and MacLeod (2017) study, the mean C-section rate was rising. Even so, they show that $\rho(\vec{x}_i)$ provides a stable ranking of C-section risk. Namely, patient i will have a higher risk of a CS than patient i' in a given year, if and only if $\rho(\vec{x}_i) > \rho(\vec{x}_{i'})$.

This risk, $\rho(\vec{x}_i)$, is estimated using the full sample and hence provides information on how variation in patient co-variates results in variation in C-section rates independent of an individual doctor's choices. Currie and MacLeod (2017) show that diagnostic skill implies greater sensitivity to the information about patient condition, $\vec{x}_i$, that is summarized in $\rho(\vec{x}_i)$. This observation can be expressed in terms of the ROC framework introduced above. Let $TPR_j = \Pr[t_{ij} = I|\alpha_i = H, j]$ and $FPR_j = \Pr[t_{ij} = I|\alpha_i = L, j]$ be the average TPR and FPR for doctor j.[17] For patient i treated by doctor j these definitions and Bayes' rule imply that the probability of intensive treatment can be written as:

$$\begin{aligned}\Pr[t_{ij} = I_i|j, \vec{x}_i] &= \Pr[t_{ij} = I|\alpha_i = H, j]\Pr[\alpha_i = H|\vec{x}_i] + \Pr[t_{ij} = I|\alpha_i = L, j]\Pr[\alpha_i = L|\vec{x}_i] \\ &= TPR_j \times \rho(\vec{x}_i) + FPR_j \times (1 - \rho(\vec{x}_i)) \\ &= (TPR_j - FPR_j) \times \rho(\vec{x}_i) + FPR_j. \end{aligned} \tag{11}$$

Letting $\theta_j = (TPR_j - FPR_j)$ and $a_j = FPR_j$, one can derive the estimation equation:

$$\Pr[t_{ij} = I|j, \vec{x}_i] = \theta_j \rho(\vec{x}_i) + a_j + \epsilon_{ij}. \tag{12}$$

The slope term, $\theta_j = TPR_j - FPR_j$ is a doctor-specific measure that increases with doctor diagnostic skill $\left(\frac{d\theta_j}{d\gamma_j} > 0\right)$.[18] Hence, doctors with better diagnostic skills are more responsive to the measure of patient appropriateness for the procedure, $\rho(\vec{x}_i)$, as long as the decision threshold, τ_j, remains fixed. Thus, θ_j provides a measure of diagnostic skill.

One can also construct a measure of procedural skill. Patients with a very high ex ante likelihood of having a C-section (e.g., $\rho(\vec{x}_i) \approx 1$), are very likely to have a C-section regardless of their doctor's diagnostic skills. Thus, one can use this subset of patients to examine the outcomes of mothers and infants following

[17] As we show above, in general these measures vary with $\vec{x}_i$. Our goal is to construct a single, one dimensional measure of skill, so we follow Chan et al. (2022) and use the mean values in this example.

[18] Proposition 5 in the appendix details this result and shows how this slope can be affected by doctor preferences. That relationship is non-monotonic; hence they are implicitly assuming that most of the variation in the slope is driven by differences in diagnostic skill rather than other factors.

C-section and attribute differences in average outcomes to the doctor's procedural skill in performing a C-sections. This point has also been used in the judge literature in which Arnold et al. (2022) use the idea of a "supremely lenient judge" to estimate the potential for misconduct by released individuals. [19] A similar computation can be done for very low-risk patients ($\rho(\vec{x}_i) \approx 0$) who are very likely to have vaginal deliveries in order to measure the doctor's skill in performing these deliveries. [20]

Thus, for each doctor j, one has estimated proxies for procedural skill and diagnostic skill. These measures can then be included as independent variables in regressions of patients' health outcomes that also include procedure prices, patient demographics, month, year, and zip code fixed effects. Two potential problems with this two-step approach are that the skill measures are estimated, and therefore measured with error, and that women may choose their physicians on the basis of their skills. To deal with these problems, Currie and MacLeod (2017) follow Kessler and McClellan (1996) and use leave-one-out, market-level averages of the skill measures as instruments for an individual doctor's own diagnostic and procedural skill measures.[21]

The identifying assumptions are that once the mother has chosen her own doctor, the skills of the other doctors in the market do not matter; that the doctor's measured skill is highly correlated with the skill of other doctors in the same market; and that the average skill level of doctors in the obstetrics health care market is exogenously determined, that is, mothers did not choose their residential locations on the basis of these measures. The inclusion of zip code fixed effects helps to control for omitted characteristics of local areas that might be correlated both with the instrument and with maternal and child health. Currie and MacLeod (2017) find that both diagnostic skill and procedural skill have significant positive effects on the outcomes for both mother and child, with the point estimates from the 2SLS model larger and more significant than the OLS estimates.

The intuition behind the model is that a doctor with lower diagnostic skill has a noisier signal of the patient's condition and is less sensitive to the appropriateness measure. A doctor with poor diagnostic skill will be less likely to correctly match the procedure to the patient: They will do more intensive procedures on inappropriate patients, and fewer intensive procedures on patients who need them. Mullainathan and Obermeyer (2022) make the same observation in the context of heart attack treatment in the emergency department. Instead of the logit model used in some of the older studies, they use a machine learning model with gradient boosted trees and LASSO to identify the patients who are good candidates for more-intensive procedures.[22] They find that doctors make systematic errors matching procedures to patients, and that these decision errors have consequences for patient survival. Like Abaluck et al. (2016), they show that this is because physicians use the wrong weights on patient characteristics when deciding on treatments—they tend to overweight a few very salient features and underweight more subtle ones. As discussed further below, this finding is consistent with a large literature demonstrating that doctors use simple heuristics based on highly salient characteristics such as patient age to make decisions and that the use of these heuristics can lead to systematic errors.

The three papers highlighted in this Section all treat doctor decision making as an information processing problem and illustrate different empirical approaches to implementing the model. Having laid out this model of decision making, the following sections use it to interpret the literature about the important factors thought

[19] Arnold et al. (2022), page 3012. They do this by extrapolating from the behavior of observed judges.

[20] Currie and MacLeod (2017) also find a positive correlation in procedural skill for both the intensive and nonintensive procedures, consistent with the hypothesis that some doctors are, on average, more skilled than others. In contrast, Chandra and Staiger (2007) hypothesize that physicians who are skilled in the intensive procedure will be less skilled in the nonintensive procedure and viceversa.

[21] There are potential limitations to this approach due to spillovers between regions (Betz et al. (2018)). Currie and MacLeod (2017) address this concern by defining markets based on the catchment area for hospitals based on patient choices.

[22] In practice one often gets the same patient risk ranking using logits obtained in more complicated AI models.

to affect it.

3 Variation in Doctor Decisions and Health Equity

A vast literature shows that doctors treat patients with similar medical conditions differently depending on the doctor's income, education, gender, and race. Appendix Table 1 outlines a number of recent correspondence studies that provide further evidence about disparities in treatment. For example, Angerer et al. (2019) sent emails on behalf of mock patients who were trying to schedule doctor appointments in Austria. They found that doctors responded more quickly and offered lower wait times to patients whose signatures indicated that they had a PhD or MD degree. Button et al. (2020) conducted an innovative correspondence study in which fictive patients sought mental health appointments. The patients randomly signaled transgender or non-binary gender identities in the text of their requests. Race was also signaled using stereotypical Black and white names. They note that mental health professionals are more likely to work in solo practices than other providers, which might give them more scope for discrimination. The results suggest some complexity in physician responses across these groups: Transgender or nonbinary (TNB) African Americans and Hispanics were 18.7% less likely to get a positive response than cisgender whites. There was no evidence of differential responses by TNB status for white patients.

As discussed below, some of these differences may be due to physician financial incentives, since higher income, or attributes correlated with higher income, could signal higher patient ability to pay. However, the evidence suggests that differences in average income are not a major part of the story. For example, Sommers et al. (2017) find that only a small fraction of reported racial differences in health care quality can be explained by the higher fraction of Black patients who lack insurance coverage, and it is not clear that eliminating financial disparities would eliminate disparities in treatment. Brekke et al. (2018) study Norwegian data in which doctors were reimbursed similarly for all patients and found that patients with more education still got longer (though fewer) visits, while less educated patients got more visits and services (such as diabetes screenings) over the course of a year. The disparities might reflect physician affinity for spending time with more educated patients, but they could also be a response to differences in time costs and health needs. Chandra and Staiger (2010) replicate the well-known finding that female and minority patients receive fewer treatments than white male patients in a sample of Medicare patients. But they also find that the health benefit of treatment conditional on detailed patient observables is lower for these patients. As they point out, "the fact that providers may offer fewer treatments to women and minorities is not by itself evidence of prejudice" since it is possible that the patients receiving fewer treatments might have fewer needs on average.[23] But if providers engage in statistical discrimination and assume that all women and minorities need fewer treatments regardless of their actual health needs, then such discrimination is problematic.

Goyal et al. (2015), Hoffman et al. (2016), and Sabin and Greenwald (2012) focus on differences in the way Black and white patients are treated for pain. Goyal et al. (2015) consider children who arrive in the emergency department (ED) with appendicitis. The underlying assumption is that most children with acute appendicitis will be treated in hospital and that the clinician they get on arrival at the ED will be approximately random. They find that Black children were less likely to receive any analgesia. Hoffman et al. (2016) explore the idea that racial disparities in treatment could be related to an erroneous belief that Black people have higher pain thresholds than white people. They find that doctors who endorse more erroneous beliefs about Black people's biological responses to pain in a survey are also more likely to

[23] Chandra and Staiger (2010), page 2.

downrate Black patients' pain when presented with patient vignettes. Similarly Sabin and Greenwald (2012) find that physicians with higher scores on an implicit bias test are less likely to say that they would give clinically appropriate oxycodone to a Black child suffering pain after bone surgery, compared to how they say they would treat a white child.

Perhaps the most popular design for studying disparities is the concordance study. The focus in these studies is on whether patients who are more similar to doctors in terms of characteristics such as race and gender receive better treatment. Cabral and Dillender (2024) obtained all Texas records for worker's compensation and for the independent medical examinations that applicants received. Assignments to doctors were random conditional on geography and the doctor's specialty. There were no effects of physician gender on the benefits received by male patients. However, female claimants seen by female doctors were 5.2 percent more likely to receive benefits. The value of benefits received was also 8.6% higher than for female claimants seen by male doctors. This finding is reminiscent of Eli et al. (2019) who study U.S. civil war veterans and show that the same physician review boards were much less likely to recommend pensions for Black veterans than for white veterans with similar medical profiles. In turn, the lower pension benefits predicted lower life expectancy for these veterans.

Some studies suggest that discordance between physician and patient characteristics can have fatal consequences (Greenwood et al. (2018); Greenwood et al. (2020); Hill et al. (2023); McDevitt and Roberts (2014); Wallis et al. (2022)). Singh and Venkataramani (2022) show that racial disparities in in-hospital mortality increase when hospitals reach full capacity, suggesting that mistakes are more likely to be made in this kind of high-stress environment and that these mistakes have the greatest impact on the most vulnerable patients.

As in Cabral and Dillender (2024), the effects are generally asymmetric: For example, Greenwood et al. (2018) find that in a matched sample, only female patients treated by male physicians are less likely to survive. Gender mismatch has no consequences for male patients treated by female physicians. Greenwood et al. (2018) find that survival increases for female heart attack patients who are being treated by male doctors in the ED when there are more female physicians present and when the doctor has treated a larger number of female patients in the previous quarter. Possibly both factors improve a male doctor's ability to interpret a female patient's symptoms.

In the case of racial discordance, Hill et al. (2023) focus on uninsured patients admitted to Florida hospitals through the ED and find that Black patients are 27% less likely to die when they have a Black physician. A nice feature of this study is that it takes the potential endogeneity of matching between patients and doctors seriously and addresses it in three ways. First, their uninsured patient pool is unlikely to have a primary care physician who can help manage their stay in the hospital. Also, admission through the ED means that these are not scheduled admissions, so the patient did not choose to arrive at a time when a particular doctor was present. Second, they develop an instrumental variables approach where the probability of concordance depends on the share of same-race physicians who are typically present during that shift (e.g. Friday nights) at the index hospital. Third, they include hospital fixed effects to account for the fact that even Black and white patients who live in the same zip code may use different hospitals.

While these correspondence and concordance studies provide compelling evidence of disparate treatment, they generally shed little light on the reasons for it. Two possible channels are either explicit or implicit biases against some groups of patients, or, more subtly, difficulties communicating across groups which could be interpreted as something that affects diagnostic skill, γ_j. Figure 4 illustrates these two alternatives. The lower curve represents a doctor with a fixed level of diagnostic skill who has different views about patients A and B. These views are represented by the slopes of the lines tangent to the curve, which, as discussed above,

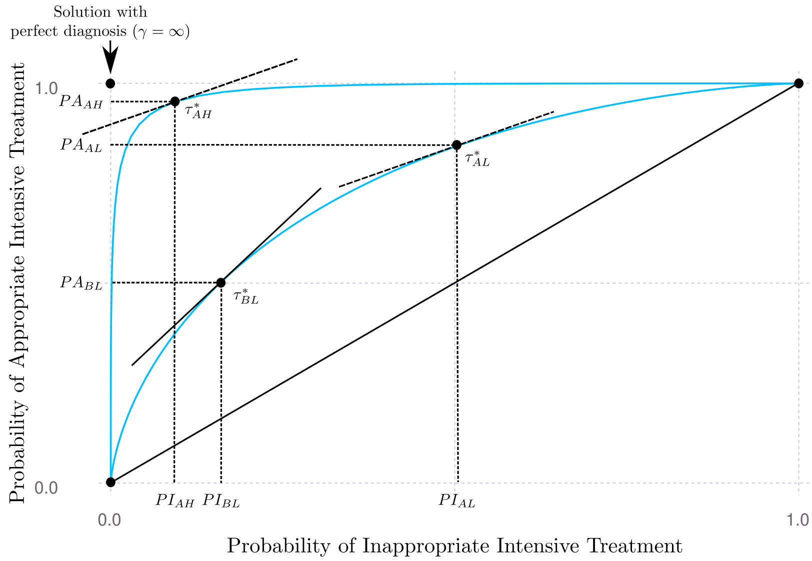

Figure 4: The Effects of Beliefs and Communication on Health Disparities

capture differences in physician beliefs about the efficacy of treatment to the two groups, and any differences in preferences for treating the two groups. As drawn, the physician is less likely to provide intensive treatment to patient B, whether it is appropriate or not. Hence, patient B will lose out on medically needed treatment when it is appropriate but may also be shielded from inappropriate treatment. An example of the latter phenomenon is that Black people were initially protected from the over-prescribing of prescription opioids at the start of the opioid epidemic by doctors' lower propensity to prescribe painkillers to them, so that the opioid epidemic was initially concentrated among white patients (Currie and Schwandt (2021))

Alternatively, suppose that the physician treating A is unable to communicate well with A, and this barrier leads the physician to choose τ^*_{AL}. In the diagram, improvements in communication would move the physician's choice of a threshold for the aggressive procedure from τ^*_{AL} to τ^*_{AH}. This change would reduce inappropriate procedure use and increase appropriate procedure use. If, for example, female doctors listen more carefully to female patients or know better what questions to ask, then this could explain the better outcomes of female patients with female doctors. In this case, the female doctor would be on the high diagnostic profile when treating female patients while the male doctor would be on the lower curve. It may also be the case that many Black patients have more trust in Black physicians, which improves communication. Lack of trust in white physicians could result from many historical injustices inflicted on Black people, including the notorious Tuskegee experiment in which Black men with syphilis were not informed of their diagnosis and were left untreated so that researchers could study the untreated course of the disease.[24]

Even doctors who make correct decisions will not be able to successfully treat patients if they cannot communicate the need for a particular course of action to the patients. Alsan et al. (2019) conduct a concordance study in which Black male patients were recruited to a special clinic offering preventive care services. They found that Black doctors were much more successful than white doctors in persuading patients to take up recommended preventive services, including diabetes screening, cholesterol screening,

[24]Alsan and Wanamaker (2018) show that this specific incident generated a legacy of distrust that endures to the present day.

and flu shots. Frakes and Gruber (2022) exploit data from the U.S. Military Health System and follow patients with severe but manageable chronic conditions, who, because of a base relocation, changed from a white to black doctor or vice versa. They find that racial concordance leads to a 15% decline in Black mortality relative to white mortality. However, only some of this difference can be attributed to differences in doctor decision making—over half of the decline is due to better patterns of medication use and adherence among patients.

Tracking down the causes of disparate treatment is important because it may help to pinpoint possible solutions. As discussed above, differences in financial resources play a role in creating disparities, so equalizing access to insurance can reduce disparities. The pain studies, and studies directly investigating physician bias, indicate that bias is an important source of disparities in care, though as Williams et al. (2019) point out, there is little evidence that interventions aimed at addressing bias have improved health.[25]

Concordance studies have concluded that the health of women and minorities could be improved by having more female practitioners and practitioners of color. For example, McDevitt and Roberts (2014) show that having even a single female urologist in a county is associated with fewer female deaths from bladder cancer. Black physicians make up only 4% of the physician workforce, so it is not possible for most Black patients to see a Black physician if they want to, or for most white physicians to have experience working alongside Black doctors. Hence, an important question for future work is whether there are additional ways to improve doctor decision making and health equity given the existing physician workforce, such as leveraging other medical professionals, including nurses or doulas, since there is greater minority representation in these fields. (Sobczak et al., 2023).

More generally, interventions that ensure that doctors correctly treat patients conditional on their symptoms can be expected to reduce health disparities. We now turn to research that measures variation in doctor decisions that arise from variation in their skill and the conditions under which they are making choices.

4 Factors that Affect the Quality of Decision Making

4.1 Skill, experience, and training

An immediate implication of the theoretical framework is that doctors with lower skill levels should set different thresholds for using intensive procedures than doctors who are more skilled. For example, Doyle et al. (2010) have an elegant study in which hospital patients were randomly assigned to the "A team" or the "B team" of residents: the A team was trained at a higher-ranked medical school. Although the two groups of patients had similar medical outcomes on average, A-team patients had systematically shorter and cheaper hospital stays. The B team used more diagnostic and testing resources to arrive at the same medical outcomes, consistent with the idea that less skilled doctors have lower thresholds for testing. In other contexts, using more resources may not be enough to compensate for lower skill. Gowrisankaran et al. (2022) find that in the Canadian province of Quebec, ED doctors with more intensive practice styles have worse patient health outcomes on average. They rely on random assignment of patients to doctors within the ED, and they measure practice style and skill as doctor fixed effects in models of procedure choice and patient health.

[25] Vela et al. (2022) conclude that the effects of most anti-bias training interventions in medical settings are either nil or extremely short-lived. They argue that this may be because the message in the anti-bias training is undermined and contradicted by other aspects of medical training. They suggest that positive interactions with both providers and patients from historically marginalized groups could have a larger impact than formal anti-bias training in terms of resetting harmful provider beliefs.

In a related context, Chan et al. (2022) suggests that since it is more costly to miss a pneumonia diagnosis than to erroneously admit a patient to hospital, less-skilled radiologists will err on the side of caution by being more likely to admit a marginal patient. They find evidence consistent with this hypothesis. Currie and Zhang (2023) also find that more skilled physicians "do more with less" in the sense of achieving the same or better health with fewer inputs.

Several studies show that doctors with more or arguably better training have better outcomes on average. For example, in models that control for hospital, quarter, day of week effects, the number of doctors present, Doyle (2020) shows that EDs have better outcomes for heart failure patients when they have a cardiologist on staff. Cardiologists have more specific training than other ED doctors, but it is possible that they are also positively selected in terms of doctor quality, so it is difficult to distinguish between selection effects and the effects of additional training per se. Schnell and Currie (2018) try to address this problem of selection versus training effects. They find that physicians from higher-ranked medical schools prescribe fewer opioids, even within the same practice address, but this could reflect either better training or the way that medical students are selected into schools of different ranks. However, they also show that in specialties that receive specific training in the use of opioids and other pain medicines, there is no difference in prescribing by medical school rank, as would be expected if doctors from higher-ranked schools were just generally better. Hence, their results suggest that training can improve practice styles.

Chan and Chen (2022) expand beyond considering doctors as providers and compare outcomes for patients treated by nurse practitioners (NPs) or doctors in Veteran's Administration Emergency Departments. They use the number of NPs who are on duty as an instrument for being treated by an NP. They find that, on average, being treated by an NP increases the length of stay and health care costs, though being treated by an NP has relatively little effect on outcomes. These results echo Doyle et al. (2010)'s finding that the "B team" uses more resources to arrive at the same results. A more striking finding is that there is considerable variation in the skill levels of both groups—many NPs achieve better outcomes at lower cost than some doctors, even though NPs have much less lengthy and intensive training than doctors.

The evidence regarding the relationship between doctor experience and outcomes is mixed. Epstein et al. (2016) focus on obstetricians and measure initial skill, defined as a physician's normalized, risk-adjusted maternal complication rate in the first year of practice. Even after 16 years, initial skill is predictive of patient health outcomes, and years of experience have little impact. Similarly, van Parys (2016) finds that the average performance of doctors treating minor injuries in an ED rises slightly with experience, but this seems to be due mainly to selection in who stays in the ED over time. Facchini (2022) estimates doctor fixed effects models and finds that obstetricians have better infant health outcomes when they have done more C-sections in the last four weeks, suggesting that it may be very recent experience that matters. Finally, Simeonova et al. (2024) evaluates the extent to which primary care physicians promote medication adherence and better health of patients on statins. Doctors whose patients do better on these measures are said to have better health management skills. However, looking at patients who had to switch doctors, they find that these skill measures appear to decay rather than to increase with a doctor's age.

One way to operationalize the idea that experience matters in the context of the theoretical framework laid out above is to make diagnostic skill and procedural skill functions of experience. For example, Currie et al. (2016) compute γ_j as described above, but allow it to vary over time. Regressing γ_j on years of experience, they find that it decreases sharply after 24 years of experience, consistent with the more negative views of the correlation between doctor experience and outcomes described above. It is possible for diagnostic skill and procedural skill to evolve in different directions with experience — a doctor might, for example, just decide to do C-sections for all patients. In this case, their diagnostic skills might atrophy while, at the same time, they became very good at performing the procedure. However, the results of Epstein et al. (2016) suggest that procedural skill, s_{tj}, is fairly flat with respect to experience, at least when it comes to doing C-sections. One difficulty with these comparisons is that we typically only observe doctors who have graduated from medical school and completed residency training, so we do not observe doctor skill levels during the period when returns to experience might be steepest.

On the whole, there has been little investigation of variation in procedural skill at the doctor level within the economics literature. Chandra and Staiger (2020) consider procedural skill at the hospital level. Arguably, while it is doctors who make decisions about how a given patient is to be treated, hospitals can influence that process. For example, a hospital can choose whether or not to have a heart catheterization facility, which will affect whether catheterizations can be performed. In terms of our framework, we can think of hospitals having a comparative advantage in either the intensive or the nonintensive procedure. Chandra and Staiger (2020) hospital's comparative advantage, such that they overuse procedures that are not their comparative advantage. In a study of the treatment of heart attack patients in 45 states between February 1994 and July 1995, they conclude that eliminating such "allocative inefficiency," that is having hospitals stick to their comparative advantage, would increase the benefits of treatment by 44%.

The papers discussed in this Section are summarized in Appendix Table 2. Overall, the research suggests that training and experience affect doctors' skill and practice styles. However, the effects of post-medical school experience seem to be small. There is also less evidence that procedural skill improves with experience than one might expect, given the well-known relationship between high surgical volumes and better surgical outcomes.[26] The evidence is also consistent with the hypothesis that selection matters, and that prospective doctors vary in their innate ability to diagnose patients and execute procedures and the extent to which they improve or keep up their skills. The empirical evidence suggests that it is unlikely that increases in the amount of training as currently practiced, or accumulation of doctor experience alone, will eliminate variations in the quality of doctor decision making.

4.2 Time pressure and fatigue

Doctors often work long hours in a fast-paced environment in which decisions must be made quickly and with little time for reflection. Time pressure could lead to mistakes if diagnostic skill, γ_j, falls with stress or fatigue. Figure 2 illustrates the idea that lowering diagnostic skill, γ_j, reduces the probability of appropriately choosing the intensive treatment and increases the probability of inappropriately choosing the intensive treatment. The more interesting point is that the increase in the use of inappropriate treatment is much greater for aggressive doctors (who move from τ^*_{AH} to τ^*_{AL}), while the decline in the probability that intensive treatments are appropriately rendered is greater for conservative doctors (who move from τ^*_{CH} to τ^*_{CL}). Hence, the same reduction in diagnostic skill has differing effects depending on the doctor's baseline type, which reflects their beliefs about the probability that an intensive treatment is likely to be appropriate and the relative efficacy of high and low-intensity procedures in their patient pool. This observation suggests that the effect of time pressures can be highly variable.

Studies focused on the impacts of time pressure and fatigue on doctor decision making are summarized in Appendix Table 3. They show a wide range of estimated effects. Tai-Seale and McGuire (2012) provide some early evidence about the importance of time pressures, showing that as the length of a visit increases,

[26] For example, Chowdhury et al. (2007) report that 74% of studies find that higher volume surgeons have better outcomes and specialist surgeons have better outcomes than general surgeons 91% of the time.

doctors are more likely to treat each new topic as the last to be covered during the visit. Subsequent authors focus on whether time pressures lead to more or less use of intensive procedures. For example, Freedman et al. (2021) find that unexpected increases in primary care physician (PCP) patient waiting times result in fewer referrals, opioid prescriptions, and Pap tests, and increases in scheduled and unscheduled follow-up visits. Persson et al. (2019) find that within an orthopedic surgeon's shift, each additional patient seen reduces the probability that a surgeon recommends surgery. On the other hand, Gruber et al. (2021) find that English ED doctors who were under pressure to reduce waiting times did so by admitting patients to the hospital, thereby increasing hospital costs by 4.9% without any effect on one-year mortality, length of stay, or the number of in-patient procedures. Similarly, Chu et al. (2024) study ED doctors and find that when doctors are managing more cases simultaneously, they order more tests, perhaps substituting testing for their time and attention.

Chan (2018) studies ED doctors and finds that as they near the end of their shifts, they are increasingly likely to admit patients to the hospital, with a 21.19% increase in the last hour of the shift, resulting in 23.12% higher costs. There are no significant effects on 30-day mortality or "bounce back" of patients to the hospital. Chan (2018) also finds that these end-of-shift effects are not found when outgoing doctors have sufficient time to hand off their patients to the incoming physician. He suggests that the changes in doctor behavior are not driven by fatigue or a higher probability of errors in judgment but by changes in doctors' valuations of their leisure time over the course of a shift. In terms of the model, δ_{Ij}, the payoff associated with the intensive procedure rises, leading to more bias in decision making.

Some of these effects of time pressure might be good for patients. For example, at the margin, fewer opioid prescriptions or orthopedic surgeries might be beneficial. But studies showing negative effects in terms of increasing the need for follow-up visits and increasing hospital costs without improving outcomes suggest that in many cases patients are harmed by time pressures.

The sign of the effect of time pressure on decisions is likely to depend on which course of action is most convenient for the doctor. In the ED, admitting the patient to the hospital may be the course of action that takes the least time, while in a PCP office, skipping tests and referrals can save time. Costa-Ramón et al. (2018) report that in a Spanish hospital, the probability that an unscheduled delivery is via C-section rises between 11:00 p.m. and 4:00 a.m. when, presumably, the obstetrician on duty would like to quickly complete the delivery and go back to bed. They note that mothers giving birth at different times of day are very similar in terms of medical characteristics that might indicate the need for a C-section.

A related question is how the doctor's emotional state impacts decision making. Chodick et al. (2023) look at the effect of a primary care doctor's encounter with a patient who has been newly diagnosed with cancer. They find a short-lived, (one hour), but large effect on the doctor's probability of ordering a wide variety of diagnostic tests, not just cancer screening tests. They discuss a number of possible reasons for this result, including a physician's emotional response to the new diagnosis for their patient or the need to test for comorbidities. Understanding the impact of a physician's emotional state, broadly defined, could help to identify moments when doctors were particularly likely to make mistakes.

4.3 The role of peers and teams

Research on the influence of peers and teams on doctor decision making has been motivated by the desire to explain geographical clusters in practice style. Proximity to peers and interactions with peers could affect physician behavior through information channels, opportunities for matching patients with physicians (or physicians with physicians), and the creation or mitigation of moral hazard within physician teams. Studies exploring these channels are reviewed in Appendix Table 4.

Several studies suggest that peers are an important source of information. For example, Agha and Molitor (2018) look at whether physical proximity to lead investigators in clinical trials for new cancer drugs is associated with faster take-up of those drugs and find that patients in the lead investigator's hospital referral region are 36% more likely to get the new drug initially, with convergence across regions after four years. Theory predicts that a doctor's threshold for using a drug or procedure is influenced by their beliefs about the proportion of patients in the population who are likely to benefit. In this case, doctors update their beliefs about whether the new drug will be beneficial for their patients more quickly when they have access to a lead investigator, or perhaps when they are more likely to see patients who have benefited from the new drug. The effects are largest in the areas with the slowest baseline rate of new drug adoption.

Chen (2021) examines patients receiving heart procedures and finds that patients do better when the surgeon has worked longer with the other hospital physicians who are caring for the patient. The effects are large: A one standard deviation increase in shared work experience reduces 30-day mortality by 10% to 14% and reduces the utilization of medical resources and length of stay. The effect is greater for more complex cases. It is interesting to compare this example to Agha and Molitor (2018) in part because it does not involve information about new or more-complex procedures. The effects presumably mainly reflect better communication among members of the team, which in turn improves patient health.

Molitor (2018) explores another dimension of peer effects—the matching of like-minded physicians in the same geographic area. Using a "movers" design, he shows that when cardiologists move to a new hospital referral region (HRR), they quickly adapt their own treatment style to the predominant style in the new region: A one percentage point increase in cardiac catheterization in the new HRR raises the doctor's own rate by 0.628%age points within one year. The effect is greater for doctors moving from low- to high-intensity areas. Since physicians do not move randomly, it is possible that the cardiologists are choosing to move to areas in which others share their desired practice style. Such sorting would increase geographic dispersion in practice styles across regions and geographic concentration in practice styles within regions.

In some situations, doctors may have little choice about how much they adopt the practice styles of others. In one of the few studies to examine the evolution of practice style during a doctor's training, Chan (2021) studies a large teaching hospital in which teams consist of junior residents who are led by a senior resident. The variation in the behavior of junior residents increases sharply after one year, when they become senior residents themselves. Medical residents presumably gain experience continuously over their first year of practice but only change their behavior discontinuously at the one year mark when they gain more autonomy. In this example, it would be wrong to attribute the junior resident's actions during the first year to their own decision making since it is apparently constrained by the senior resident.

Silver (2021) focuses on teams of ED doctors and exploits variations in the composition of teams across shifts, arguing that these are essentially random. He finds that doctors work faster when they are placed with a fast-paced team and that, on average, the faster pace has no effect on the outcomes of discharged patients. However, the riskiest patients suffer increases in 30-day mortality. This result contrasts with Gruber et al. (2021) who, as discussed above, find that physicians working faster in response to a mandate to reduce ED wait times increased costs, without having any negative effects on patient health. Possibly, the American doctors were under greater pressure not to increase costs than the British doctors in Gruber et al. (2021), but the contrasting results suggest caution when extrapolating from any one study in this doctor peer effects literature.

While Silver (2021) and Gruber et al. (2021) suggest that doctors can choose to work faster or slower, Chan (2016) asks whether doctors who work more slowly are shirking and thereby forcing other members of their team to work harder. His study focuses on two teams working in the same hospital. In the first team, doctors decided how patients were allocated within their group. In the second team, patients were initially assigned to doctors by a nurse scheduler, and then the regime changed so that patients were assigned by the doctors themselves. Chan (2016) shows that switching the nurse-managed team to being doctor-managed reduced wait times by 13.67 percent without any effects on costs, utilization, or health. His interpretation is that doctors shirked under the nurse managers, but that doctor-managers had a better understanding of how long each patient should take, so that they were better able to detect and prevent shirking. The authors discount the alternative explanation that supervising doctors are better able to match patients to the doctors because there was no change in health outcomes.

Currie et al. (2024) examine peer effects in physician prescribing to adolescents with mental health conditions. They point out that it can be difficult to identify peer effects if doctors with similar training and experience tend to have practice styles that evolve similarly over time and also cluster in the same locations. They look at correlations between the index physician's probability of prescribing inappropriately, the probability that physicians with similar training and experience from outside the area prescribe inappropriately, and the probability that physicians from the same area but with different training and experience prescribe inappropriately. They find that the "effect" of physicians from the same cohort but outside the area is about half the size of the effect of local physicians from different cohorts. Hence, some of what appears to be a peer effect actually reflects the co-evolution of practice styles among similar physicians. The size of the spillover effects are consistently larger for non-psychiatrists than for psychiatrists, indicating that specific training can mitigate the extent to which inappropriate prescribing than peer effects are the most important determinant of variations in practice style in their data.

These papers suggest that it is quite difficult to identify true peer effects outside of certain specialized settings in which it is plausible to assume that doctors do not choose their peers. Hence, we are a long way from being able to use estimates of peer effects to think about influencing doctor behavior.

4.4 Financial incentives

Health economists have long realized that doctors can be influenced by financial incentives. Handel and Ho (2021)'s chapter in the Handbook of Industrial Organization provides a review of some aspects of the healthcare market that impact doctors' financial incentives, including competition in hospital and insurance markets, negotiations between hospitals and insurers, and increasing vertical integration in hospital markets.[27] In our model, the δ_{tj} parameter captures the pecuniary returns that doctor j receives from choosing procedure t. Appendix Table 5 provides an overview of some post-2010 contributions to the large literature on financial incentives in health care markets. While the findings of some studies can be characterized by an estimated elasticity, in many cases that is not possible because the financial changes in question are very lumpy (such as moving from fee-for-service to capitated payments) or may involve non financial transactions as well as the purely financial, as in the case of drug detailing. Two overarching questions addressed in this Section are whether and how governments and insurance plans can use financial incentives to reduce health care spending without worsening patient health and whether some types of patients are more or less

[27] The IO literature they survey has focused on the larger players, such as hospitals and insurers which can be understood as "firms," rather than on the decisions of individual physician providers. However, as more physicians work for large groups, and more practices become part of vertically integrated health care companies, this distinction may become less relevant. For example, Chernew et al. (2021) show that vertically-integrated physicians increase inpatient hospital care for elderly patients rather than substituting for it.

vulnerable to the distortions in doctor decision making that are induced by financial incentives.

Several studies look at changes in reimbursements from the U.S. Medicare program. Reducing spending in Medicare is of particular interest both to policy makers and economists as the population ages and advances in medical technology make Medicare spending an increasing part of the federal budget.[28] Clemens and Gottlieb (2014) take advantage of a consolidation of Medicare reimbursement regions that raised reimbursements in some areas and lowered it in others. They show that higher reimbursement rates increased the use of elective procedures and the probability of hospitalization for heart attacks (acute myocardial infarction) within one year, without having any effect on four-year mortality rates. The elasticities are greater than one, suggesting that the supply of elective procedures is very responsive to prices. Note that if hospitalizations were driven primarily by consumer demand, higher prices would lead to lower quantities. Hence these results suggest that the marginal hospitalization is driven by supply-side considerations.

A major complaint about Medicaid, the U.S. public health insurance program for low-income individuals, is that it is difficult for patients to get an appointment, and one reason for this may be that Medicaid payments are much lower than either private health insurance or Medicare payments. Bisgaier and Rhodes (2011) report on an audit study in which patients on Medicaid were six times more likely to be denied a specialist appointment than patients with private health insurance and had to wait three weeks longer to see a provider if they did get an appointment. The implied elasticity of visit availability with respect to payments was 2.65. Alexander and Schnell (2024) look at a Medicaid "fee bump" that resulted from the 2010 Affordable Care Act (ACA). The ACA gave payments to states to try to reduce the payment gap between Medicaid and other payers. The fee bump increased Medicaid payments by an average of 60%, with considerable variation across states. Their estimates suggest that closing the gap between the payments offered by Medicaid and those offered by private health insurance would eliminate disparities in access to primary care for children and would also reduce access disparities by two-thirds for adults. Similarly, Cabral et al. (2021) study a Medicare reform that increased provider payments and estimate that it increased provision of targeted services by 6.3% with an elasticity of services to payments of 1.2. Dunn et al. (2024) consider another type of provider disincentive associated with Medicaid — an elevated risk of having a claim denied or otherwise unpaid. They find that 18% of Medicaid claims are denied, a much higher rate than under either Medicare or private insurance. They conclude that this high probability of non-payment is as great a barrier to doctors accepting Medicaid patients as the lower fees.

Other authors focus on the effect of capitation—that is, providing doctors with a fixed payment per patient. Most economists would predict that capitation would lower the intensity of service delivery relative to fee-for-service payment, which is exactly what empirical studies have found in empirical studies. For example Ding and Liu (2021) show that providers with capitated payments used 12.2% fewer resources (especially physical therapy and diagnostic testing) compared to non-capitated providers, with no change in outcomes. One issue with studies of capitation is that providers who are not being reimbursed for providing specific services may have little incentive to record them in claims data. Hence, some of the measured reduction in services rendered could be an artifact of changes in reporting practices.

Chorniy et al. (2018) show that doctor behavior can be affected by the specific incentives built into managed care contracts. In their South Carolina setting, Medicaid providers who were switched to capitated payments plans from fee-for-service plans got larger payments if patients had specific chronic conditions. Providers were also penalized if they screened children for chronic conditions at lower than average rates. Chorniy et al. (2018) follow the same children over time as their providers were switched from fee-for-service

[28]Medicare accounted for 12% of the total federal budget in 2022. See https://www.pgpf.org/budget-basics/medicare.

to capitated contracts. They find an 11.6% increase in diagnoses of ADHD and an 8.2% increase in diagnoses of asthma without any effect on ED use or hospitalizations. These findings suggest that more research looking at specific compensation contracts for doctors is warranted.

Several more tailored schemes for reducing health care costs without reducing quality have also been evaluated. Alexander (2020) studies a New Jersey policy that allowed hospitals to select into a program that offered physicians incentives if they lowered the costs of care. Alexander (2020) finds that the program had no effect on costs or procedure use—instead, physicians were able to game the system by directing their lowest-cost patients to participating hospitals. This simple tactic lowered patient costs at these specific hospitals so that doctors could reap the incentive payments. This behavior resulted in higher patient travel costs.[29] Alexander and Currie (2017) show that doctors' responses to incentives may also be affected by factors such as capacity constraints. They find that doctors are generally more likely to admit child respiratory patients when those patients have private insurance rather than lower-paying public insurance. This gap grows when beds are in high demand because of high flu caseloads.

Strong responses to physician financial incentives have also been found in European settings, where most countries have some form of universal health insurance coverage. For example, Wilding et al. (2022) focus on an English policy that imposed financial penalties on general practitioners (GPs) when the fraction of hypertensive patients with blood pressure under control fell below a target. They show that stricter targets increased prescription of anti-hypertensive medication. But doctors also showed evidence consistent with gaming: They did multiple tests on patients whose blood pressure initially exceeded the threshold (presumably trying to get a reading below the threshold), took actions to have patients declared exempt from testing requirements, and were more likely to report that patients exactly met the threshold, suggesting greater use of rounding. In France, Coudin et al. (2015) show that the imposition of price controls increased the number of procedures by over 80%, suggesting that physicians increased quantities to make up for shortfalls in income due to the price controls.

As we pointed out at the start of this section, it might be more surprising to economists to find instances in which doctors did not respond to financial incentives. Some recent studies focus on factors that mute or mediate the expected relationship, including a variety of patient characteristics. For example, Johnson and Rehavi (2016) look at patients who are themselves physicians. They find that physician patients are about 6% less likely than other well-educated patients to have unscheduled C-sections, and that financial incentives affect C-section rates only for nonphysician patients. However, it is not entirely clear whether this null result reflects pushback from informed consumers or physicians refraining from suggesting unnecessary C-sections to their peers.

Chen and Lakdawalla (2019) use the same change in Medicare billing areas as Clemens and Gottlieb (2014) and ask how physician responses to changes in Medicare reimbursements vary with the income of the patient. A key institutional detail is that fee-for-service Medicare patients have copays. Since richer patients are likely to have a greater willingness to pay than poorer ones, the authors predict that higher reimbursements will lead to larger increases in procedure use in richer patients because poorer patients are more likely to resist the higher copays. They show that increases in reimbursements increased the gap in services received between high- and low-income patients, implying that the supply of services is increasingly

[29] In contrast, Gupta (2021) studies the impact of the Hospital Readmissions Reduction Program (HRRP) which applied to all hospitals and penalized hospitals with Medicare readmission rates that were higher than a given threshold. He finds very large effects of the program: the HRRP was estimated to account for two-thirds of the observed reduction in readmission probabilities and to have reduced 1-year mortality by 8.87%. These positive effects were achieved by increasing the intensity of care during the initial hospital admission. The contrast between these two papers shows that details, such as whether the policy applies to all hospitals or a subset, matter.

elastic as patient income increases.

Whether the physician has an ongoing relationship with a patient has also been shown to be an important mediator of the extent to which financial incentives affect patient care. Brekke et al. (2019) use Norwegian administrative data linking health, national insurance, and labor market participation to examine physician behavior with respect to the issuance of sick-leave certificates. In order for workers to claim sick-leave benefits, they must have a doctor sign a certificate. Physicians see patients both in their own practices and in EDs. They are likely to have ongoing relationships with patients in their own practices but not with patients in the ED. Physicians may also be on fee-for-service or fixed-salary contracts. The authors show that physicians are 34.63 percent more likely to issue sickness certificates for their own patients with fee-for-service contracts and 24.15% more likely with fixed-salary contracts. However, for new GPs with fixed salaries, there is no gap in rates between own patients and ED patients, which may reflect the fact that new GPs do not yet have any ongoing relationships with patients. The size of the gap in sick leave issuance between own patients and ED patients is greater in areas with larger numbers of GPs per capita and among GPs who have openings for new patients, suggesting that competitive pressures also influence this behavior.

Currie et al. (2023) also examine the impact of competition on physicians, using state laws that allowed nurse practitioners to prescribe controlled substances independently as a source of exogenous variation in competition. They find that general practitioners responded by prescribing significantly more controlled anti-anxiety medications, more opioids, and more co-prescriptions of the two types of drugs. The impact of the change in laws was greater in areas with higher ratios of NPs per GP to begin with and was concentrated in specialties that face the most competition from NPs. Their findings suggest that in some cases, competition can have harmful effects on patients by leading to the over-provision of services.

We will briefly touch on two other types of physician incentives here, those due to "detailing" and those due to malpractice. Detailing is the practice of marketing drugs and other medical equipment or products directly to physicians. In some cases, this may involve visits from company representatives providing information, but often detailing also involves a payment to the physician in cash or in kind (e.g., meals or travel expenses). U.S. sunshine laws passed as part of the 2010 Affordable Care Act require companies selling pharmaceuticals and medical devices to report most payments made to physicians to the federal government.[30] These disclosures have enabled researchers to learn more about these payments and their impacts on physician behavior. Carey et al. (2021) examine the impact of detailing on the use of generics and the efficacy of drugs prescribed. They find that the size of payments does not matter much. Even a small payment increases prescribing of the detailed drug by about 2% in the six months following receipt of a payment. However, doctors do not seem to be prescribing less-effective drugs or delaying transitions to generics.

Shapiro (2018) also suggests that the effects of detailing are relatively benign. He studies an antipsychotic drug, Seroquel. Two clinical trials showed that Seroquel had a better side-effect profile than leading competitors. Building on early work by Azoulay (2002) that suggested that the impact of drug research is amplified by marketing, Shapiro finds that these trials had little impact on prescribing unless they were accompanied by detailing visits. He interprets this as evidence that the new information from the trials was conveyed to doctors through detailing. Detailing visits after the trials resulted in small shifts in prescribing towards Seroquel, and more of these prescriptions were "on-label," that is, for indications approved by the U.S. Food and Drug Administration (FDA).

In contrast to Carey et al. (2021) and Shapiro (2018), Newham and Valente (2024) find that payments

[30]In response to the 2018 U.S. SUPPORT Act, CMS Open Payments started including payments to physician assistants, nurse practioners, clinical nurse specialists, certified registered nurse anesthetists, anesthesiologist assistants, and certified nurse-midwives. Additional research is needed to study the effects of this expansion of reporting requirements.

to physicians increase prescribing of branded rather than generic diabetes drugs, raising costs. Carey et al. (2024) also find that marketing payments increase expenditures on cancer drugs in Medicare with no subsequent improvement in patient mortality. As more years of CMS Open Payments data become available, further research will be possible to help clarify this issue, though the existence of these data may itself shape the course of pharmaceutical marketing in the years to come.

Agha and Zeltzer (2022) extend the peer effects literature discussed above to consider the impact of detailing on physicians who do not receive payments directly but who share patients with doctors who received payments. Using Medicare claims data, they find that such spillovers account for a quarter of the increased prescribing that results from detailing payments. The effects are larger for physicians who share more patients with the doctor who received drug company payments. This finding is particularly important in that it underscores the limitations of sunshine laws in tracking the influence of pharmaceutical companies on physicians.

Doctors themselves often cite fear of malpractice as a factor that influences them to practice defensive medicine—that is, the practice of ordering unnecessary procedures and tests to protect against malpractice risk. In practice, the risk of financial loss is mitigated by malpractice insurance. And since malpractice insurance is not experience rated, doctors typically do not even face higher insurance premiums after a finding of malpractice. Hence, it may be the unpleasantness associated with being sued and the subsequent damage to their reputations that doctors wish to avoid rather than financial penalties per se.

A large literature leverages changes in state laws to assess the impact of malpractice on doctor behavior. Mello et al. (2020) offer a survey of this literature and conclude that while some authors find non-zero effects, the impacts of changes in laws governing malpractice are typically quite small. Nevertheless, the National Academy of Sciences (Balogh et al. (2015)) notes that the malpractice system could have a negative systemic effect by inhibiting reporting of, and learning from, diagnostic errors.

Currie and MacLeod (2008) offer several possible reasons for the small estimated effects of malpractice reforms. First, most studies lump all changes in tort laws together, even though different types of laws are predicted to have different effects. For example, laws capping damages may encourage reckless behavior while reforms making physicians liable for the share of the damages that they caused (rather than allowing plaintiffs to sue the "deep pocket" in the case for 100% of damages)[31] should have the opposite effect. Second, the impact of a law change is likely to depend on whether a physician is doing too many or too few intensive procedures to begin with. For example, if a doctor was causing harm by doing unnecessary C-sections, then raising the cap on damages (for example) might cause them to reduce the number of C-sections. On the other hand, if a doctor was doing too few C-sections, then the same law change might cause them to do more. Frakes (2013) captures this intuition. The key question in most malpractice cases is whether the doctor provided care consistent with accepted medical practice. As of the late 1970s, most states used state standards to define accepted practice. But over time, many states moved to using national rather than state-level norms. Frakes (2013) shows that state C-section rates tended to converge to the national rate after this change, with no change in infant health.

In summary, recent work adds to voluminous existing evidence that physicians respond to financial incentives. But it goes further by showing how difficult it has been to use this fact to either rein in health care costs or improve the quality of care. Doctors are not unique in being difficult to properly incentivize through

[31] Joint and several liability makes a defendant liable for the full harm suffered by a plaintiff even if the defendant is only responsible for a small portion of the harm. Many U.S. states have reformed their tort laws in ways that try to limit each defendant's liability to the share of the damages that they caused or that shield defendants who are responsible for only a small fraction of the harm from being sued for the full amount.

the price system. In professions where there is a noisy relationship between inputs and outputs, tinkering with input prices or rewarding/penalizing outcomes is unlikely to elicit socially optimal performance. It is important to actually measure and reward the appropriateness of the inputs and their contribution to the observed outcomes. Doctors often respond to changes in reimbursement rates by changing diagnoses or recommending additional services and may respond to penalties by avoiding certain patients or over- or under-providing services. Hence, manipulation of the price system can have many unintended consequences. Research asking which types of patients are most affected by the unintended consequences of changes in financial incentives has provided some initial answers suggesting that less-educated and lower-income patients who lack medical homes are most impacted, but this is an interesting question for further research. Research into other changes in financial incentives such as those from detailing payments or threats of malpractice have so far suggested relatively mild effects on physician behavior, though large changes, such as drastically weakening the threat of malpractice, might have larger effects.

The existence of financial incentives moves utility-maximizing doctors away from what an altruistic doctor would do. But since providing medical care has both social costs and benefits, altruistic doctors who care only about patients may provide too much care Chandra and Skinner (2012). Adding fee-for-service payments would cause doctors to provide even more care, while a capitated system incentivizes less care. How far the care actually provided under different payment schemes is from a socially optimal level of care is an open but difficult question that would require grappling with the social value of health. Another interesting question is how much money doctors leave on the table because they are altruistic (and/or care about their reputations). Studies on responses to financial incentives imply a wide range of response elasticities. It would be useful to study how these elasticities are related to doctor, patient, procedure, and market characteristics.

5 Improving the Quality of Doctor Decision Making

So far this survey has demonstrated that there is a great deal of variation in the quality of doctor decision making and that poor decisions can have a negative effect on patient health, increase health care costs, and widen health disparities. There is a growing literature discussing possible ways to improve doctor decision making beyond adjusting payment systems. This Section discusses research considering the effectiveness of providing information to doctors and/or patients, using heuristics or guidelines, or using new technologies, such as electronic medical records and decision support tools, in an attempt to improve medical decision making. We can think about these is in terms of whether they 1) target diagnosis (γ_j); 2) whether they try to shift the doctor's priors regarding the usefulness of a medical procedure for the two types of patients, $\Delta_{LNIj}/\Delta_{HIj}$; or 3) whether they affect the doctor's beliefs about the relative proportions of patient types, p_{Lj}/p_{Hj} in the population. At the extreme (e.g. guidelines that specify or proscribe particular actions in specific cases), they might involve taking decision making out of the doctor's hands.

5.1 Providing information

A number of studies explore the consequences of providing information about practice style to either physicians, patients, or both. Appendix Table 6 summarizes several examples from this literature. The most straightforward studies are experiments in which letters were sent to randomly selected treatment physicians; control physicians did not receive letters. For example, Sacarny et al. (2016) designed a randomized controlled trial targeting physicians who were high prescribers of Schedule II controlled substances (opioids, amphetamines, and barbiturates) to Medicare patients. This intervention could be interpreted as an attempt to reach doctors who were consistently over-estimating the share of patients in their practices who were likely to benefit from these drugs. If these doctors can be persuaded to raise their estimate of the relative proportion of patient types p_{Lj}/p_{Hj} in their patient pool, then this would cause them to raise their threshold for prescribing, τ_j. Doctors in the treatment group received letters informing them that their prescribing patterns deviated significantly from those of their peers. These letters resembled comparative billing reports that Medicare routinely sends to providers comparing their billing practices to those of their peers and did not mention any sanctions. Regarding results, the title of the paper says it all: "Medicare Letters To Curb Overprescribing Of Controlled Substances Had No Detectable Effect On Providers." There was no evidence of heterogeneous effects by prescriber specialty, region, or whether the prescriber had been investigated for fraud.

However, several subsequent studies have found significant effects of similar letters on physician prescribing. In a follow-up paper, Sacarny et al. (2018) targeted outlier prescribers of the antipsychotic drug quetiapine and sent them three letters highlighting their outlier status relative to peers. Over the nine months of the experiment, the number of days of quetiapine prescribed fell by 11.1 percent in the treatment group relative to the control mean, and the reduction lasted at least two years. This reduction was largest for patients with low-value indications, and there were no negative effects on patient health. It is possible that receiving three letters over a short period made the intervention seem less like a routine "form letter" and more like there was an implied threat of some sort of sanction.

Ahomäki et al. (2020) report that a precautionary letter sent to Finnish physicians who were prescribing high numbers of paracetamol-codeine pills to new patients reduced the number of pills prescribed to new patients by 12.8% of the treatment group baseline, which is similar to the more recent Sacarny et al. (2018) paper. Again, the letter may have carried an implicit threat, since such letters are not routine in the Finnish context. Hence, the question raised by these papers is whether doctors are responding to the information contained in the letter, or whether they are afraid of being sanctioned for their outlier behavior. Possibly the important information being conveyed is not so much that they are outliers, but that an authority is watching their prescribing behavior.

In perhaps the most famous recent example of a letter-writing intervention, Doctor et al. (2018) started with vital statistics mortality data from California identifying people who had died from overdoses of prescription opioids. Then, using the state's prescription drug monitoring program (PDMP) records, they located the doctors who had prescribed the fatal drugs. The experimental intervention involved sending a letter to a treatment group drawn from these doctors informing them that their patient had died of an opioid overdose. The researchers could then monitor the treatment doctors' subsequent opioid prescribing using the PDMP. They found a 9.7% reduction in the prescribing of morphine equivalent milligrams of opioids in the three months following the intervention. Of the "letter experiments" discussed here, this one arguably comes closest to a pure information intervention. The researchers were not writing on behalf of any state or regulatory agency, so there was less of an implicit threat. And they were supplying information that doctors would not necessarily be able to acquire easily from other sources—when U.S. doctors treat a patient who does not return, they are not routinely informed about whether this is because the patient moved, switched physicians, stopped going to the doctor, or died.

A second group of "informational" studies seeks to measure the effect of new clinical knowledge on physician behavior. For example, in a meta-analysis, Hammad et al. (2006) suggested that selective serotonin re-uptake inhibitors (SSRIs) increased suicidal thinking in children and young adults. A preliminary version of this study led the FDA to put a prominent warning label on SSRI drugs in 2004. Early studies such as Gibbons et al. (2007) indicate that these warnings led to sharp drops in prescribing to children and adolescents in the U.S. and Norway, and to declines in prescriptions of SSRIs generally. Building on this evidence, Dubois and Tunçel (2021) replicate the finding in French data and then build a random coefficient discrete choice logit model to examine changes in physician prescribing across several drug classes. They find reductions not only in SSRIs but in the prescribing of close substitutes, and an increase in the off-label use of other types of psychiatric drugs as treatments for depression. A quarter of physicians stopped prescribing SSRIs altogether, but considerable variation in physician prescribing remained both before and after the change. A limitation of their work is that their model must perforce rely on the strong assumption that the way physicians are matched to patients does not change following the announcement.

McKibbin (2023) presents another convincing study of the impact of new information. Since FDA approval is a lengthy process, many sick cancer patients do not have time to wait for the process to be completed but take promising new drugs "off label" (i.e., before they are FDA approved for that indication). McKibbin (2023) looks at what happens to off-label use of cancer drugs when new drug trial information becomes available. She finds that physician responses are sensitive to whether the p-value is less than 0.05. When the effect of the drug is deemed statistically significant, demand doubles in the year after the finding. If the drug is found not to have a significant effect, demand falls by a third over the next two years. Avdic et al. (2024) also find asymmetric responses to new information. Their study focuses on drug-eluting stents used in heart surgery. The new stents were first thought to be an improvement and then shown to be inferior to older stents. Using Swedish data, Avdic et al. (2024) show that doctors were slow to take up the new stents but abandoned them quickly when the new information about their potentially harmful side effects came out. DeCicca et al. (2024) examine the effect of a prominent study that showed that C-sections were unnecessary for breech birth and show that doctors rapidly reduced the frequency of C-sections for breech babies at a time when C-sections were rising rapidly. These studies suggest that understanding how physicians respond to new information is an important question for future research.

Howard and Hockenberry (2019) ask how the uptake of new information from clinical studies is affected by physician age. The specific example is new information about episiotomies from clinical studies showing that they are ineffective in reducing complications of labor and delivery. They find that doctors with over 10 years of experience were much less likely to change their practice in response to the new information. However, they also find that the gap between new and old doctors was smaller in teaching hospitals, which are more likely to promote the adoption of evidence-based medical practices.

Wu and David (2022) provide an example that fits nicely into the theoretical framework laid out above. They consider the choice of minimally invasive versus "open" surgical procedures for hysterectomy. In 2014 the FDA announced that the minimally invasive procedure had a previously unappreciated risk of spreading a rare form of cancer. This announcement changed the expected benefit of the intensive relative to the non-intensive procedure ($\Delta_{LNIj}/\Delta_{HIj}$). But the authors point out that this ratio also depends on the surgeon's relative skill in performing the two procedures. While overall use of the minimally invasive procedure fell, it actually rose among the subset of surgeons who were much better at performing the minimally invasive procedure than the open procedure.

Together with the "letter experiments" discussed above, these studies suggest that doctors pay more attention to some types of new information than others and that the impact of new information can vary with characteristics such as experience and skill. An important question going forward is what factors make information salient and whether these factors predictably vary with other physician characteristics in a predictable way.

Information provided to both physicians and consumers in forms such as “quality report cards” can also influence physicians. Kolstad (2013) considers two potentially important effects of the introduction of new report cards for coronary artery bypass graft (CABG) surgery. Report cards create an “extrinsic” incentive for surgeons to improve their scores in order not to lose business. But knowing how they are doing relative to other surgeons may also spur physicians to improve their practices for the “intrinsic” reason that they get utility from improving patient’s health and realize that they could be doing better. Kolstad (2013) estimates a structural model of consumer demand in order to separate intrinsic from extrinsic motivations. Improvements made in response to predicted changes in consumer demand are thought to reflect extrinsic motivation, while the remaining change in doctor behavior after report cards are introduced is defined as change due to intrinsic motivation. He finds that intrinsic motivation is more important than extrinsic considerations and that the response to report cards is greatest for physicians who are revealed to be worse than other surgeons in their own hospitals. This last finding opens the door for a third type of motivation—surgeons who are worse than other surgeons in their own hospital fear loss of business or penalties for poor performance. Alternatively, physicians may perceive other physicians in their own hospitals as a more relevant comparison group than other physicians.

Finally, one can ask how extraneous information affects doctor decision-making. Persson et al. (2021) focus on children who have a higher probability of being diagnosed with ADHD simply because they are “young-for-grade.”[32] They show that the “extra” diagnoses induced by being young for grade cause a child’s siblings to also be more likely to be diagnosed with ADHD. While some part of this increase is likely due to an increase in the probability that children are presented for evaluation, due to choices made by parents or schools, it is ultimately up to the doctor to make a diagnosis or prescribe medication. Hence, this example suggests that doctors’ decisions can be influenced by erroneous information about siblings. Similarly, Ly et al. (2023) find that giving doctors charts saying a patient has congestive heart failure makes them less likely to test for pulmonary embolisms, regardless of the other features of the case.

In sum, the research discussed in this Section shows that information provision can impact practice style. However, information provision does not eliminate undesirable variations in practice and does not always even lead to changes in the right direction. In terms of the model, this result may suggest that inaccurate beliefs about the usefulness of a medical procedure (or drug) for the two types of patients, $\Delta_{LNIj}/\Delta_{HIj}$; or about the relative proportions of patient types, p_{Lj}/p_{Hj}, may not be a main driver of improper care. In view of the fact that a “helicopter drop” of information does not always have the desired effect, we next consider the role of various types of heuristics and guidelines.

5.2 Heuristics and guidelines

Simon (1957) introduced the idea that because people are boundedly rational, they often take mental short-cuts and apply simple rules as aids in decision making. The properties of these rules, or heuristics, were further explored by Daniel Kahneman and Amos Tversky in many works (but see especially, Kahneman et al. (1982)). Heuristics are powerful because they often work well, though following them can also lead to systematic errors. We will use the term “guideline” to denote something more formal than a heuristic in that it is a set of rules laid down by an authority such as a professional association or a government agency. Guidelines usually do not have the force of law and there are typically few or no penalties for violating them,

[32]Since ADHD is a neurodevelopmental condition that is usually present from birth, small differences in children’s birthdates should not affect the underlying probability of having ADHD. Yet children born right before school entry cutoffs, who are therefore “young-for-grade,” have been shown to be more likely to be diagnosed with ADHD.

but they do provide clear expectations about appropriate (or inappropriate) behavior. Appendix Table 7 provides an overview of studies that address two questions. First, do doctors follow simple heuristic rules and if so, what effect does this have on patient health care utilization, costs, and health? Second, can diagnostic skill, γ_j and patient health be improved by physician adherence to guidelines?

The use of simple decision rules is a ubiquitous human behavior, so it would be surprising if doctors did not use them. Health economists have shown convincing evidence that heuristics not only exist in medicine but can have important consequences for patient health care utilization, costs, and health. In an ingenious paper, Almond et al. (2010) look at the treatment of newborns with birth weights on either side of a 1500 gram threshold that is used to define "very low birth weight." They show that infants just below the threshold receive more medical care and are more likely to survive than infants just above the threshold. This result suggests that many infants above the threshold are erroneously denied the care that could save them because of a too literal adherence to the decision rule implied by the 1500 gram cutoff. Infants around the 1500 gram cutoff may be more or less sick depending on additional factors such as lung development. Closer attention to other indicators, in addition to birth weight, could improve the targeting of care.[33]

Geiger et al. (2021) use a similar regression discontinuity design to examine the effect of a designation of "advanced maternal age" (AMA) for pregnant women who will be aged 35 or more on their expected delivery dates. They find that AMA mothers receive more screening and specialty visits and that this additional care has a large effect on perinatal mortality (infant death in the first month). As in Almond et al. (2010), this result suggests that rigid reliance on a simple heuristic based only on maternal age harms some patients who would have benefited from more care. The effects are greatest for pregnancies without obvious risk factors, suggesting that many apparently low-risk women would have to be more intensively screened and treated in order to prevent the marginal deaths.

Currie et al. (2016) find that doctors treating heart attack patients in Florida also appear to rely on age to ration treatment. They are less likely to treat older patients aggressively, even though they estimate that all patients would have benefited from aggressive treatment in terms of a reduced risk of hospital readmission and mortality. Olenski et al. (2020) look more specifically at coronary artery bypass graft surgery (CABG) for heart patients using a regression discontinuity around a patient's 80th birthday. They find that patients admitted in the two weeks after their birthday are 28 percent less likely to receive CABG than patients admitted in the two weeks before. Coussens (2018) uses a regression discontinuity design to see whether the probability of being tested, diagnosed, or admitted for ischemic heart disease is higher when a patient is over age 40. The results suggest that testing increases almost 10% at age 40, while diagnoses and admissions increase by 20%. Effects are larger for patients presenting without chest pain and for female patients, who are less likely to experience the stereotypical symptoms of heart disease. One might expect doctors to be more likely to use heuristics when they were busy but Coussens (2018) finds the reverse—the effect of the age threshold is larger when the ED is less busy and in the first half of the doctor's shift. Geiger et al. (2021), Olenski et al. (2020), and Coussens (2018) all highlight that physicians have a tendency to "think discretely" about continuous patient characteristics such as age.

These articles provide strong evidence that doctors use simple heuristic cutoffs for providing care and that they do not necessarily assess each patient individually on the merits of their cases. Moreover, these decisions matter for patient health. However, this observation does not necessarily imply that heuristics are undesirable or inefficient. Only in a world with unlimited time and resources would we not want (or need)

[33] Barreca et al. (2011) show that the regression discontinuity design employed by Almond et al. (2010) is sensitive to measurement error (heaping) in birth weights at the threshold. However, Almond et al. (2011) show that their main results are robust to the use of a "doughnut" design that excludes observations that are very close to the threshold.

to use them. An important question then is whether these simple rules could be enriched in a way that meaningfully improves doctors' choices and patient health without greatly increasing health care costs.

Guidelines tend to be more complex than simple heuristics and may be especially helpful for decisions that do not involve a simple zero-one choice. For example, Currie and MacLeod (2020) consider guidelines for drug treatment of adult depression. There are many treatment choices, and it is not possible to know a priori which drug is best for a particular patient. There may be a trade-off between choosing the drug with the highest expected value and experimenting to find a drug that may be better for a particular patient. The downside of experimentation is that it can expose patients to the risk of poor outcomes because many drugs have side effects. A novel implication of their model is that experimentation is only useful if the doctor has enough diagnostic skill to learn from it and is willing to change their underlying beliefs about the efficacy of the treatment. Using claims data, they show that patients of more-skillful doctors (psychiatrists) benefit from experimentation, while patients of less-skillful doctors (GPs treating mental illness) derive little benefit from experimentation. The model predicts that higher diagnostic skill leads to greater diversity in drug choices across patients and better matching of drugs to patients even among doctors with the same initial beliefs regarding drug effectiveness. They also show that conditional on doctor skill, increasing the number of drug choices predicts poorer patient health by making it more likely that the doctor will choose a drug that is a bad match.

Can the use of guidelines improve outcomes? Medical guidelines vary from being very prescriptive (e.g., all heart failure patients should get beta blockers unless there are contraindications) to being rather loose and aimed not at mapping specific actions to specific conditions but at eliminating harmful choices. For example, a guideline might recommend that doctors avoid prescribing multiple psychiatric drugs at the same time without specifying which drugs they can use. Guidelines may come from government agencies (as in the case of the English National Institute for Health and Care Excellence) or from professional associations such as the American Psychiatric Association. As in the case of heuristics, guidelines are usually not compulsory though physicians who violate guidelines could in some cases expose themselves to legal liability. Currie and MacLeod (2020) explore the rather loose guidelines that the American Psychiatric Association has drafted for adult depression treatment. These guidelines focus on changing drugs when an initial drug is found to be ineffective and on the inadvisability of prescribing multiple drugs at the same time. They show that the patients of physicians who violate these guidelines have significantly worse outcomes than other patients.

Cuddy and Currie (2020) focus on guidelines for treatment of adolescent depression and anxiety. These guidelines are considerably more detailed and more prescriptive than those governing treatment of adults. Using claims data, they show that guideline violations are widespread. Cuddy and Currie (2024) build on this work by showing that these guideline violations are consequential. In order to deal with the possibility that patients are demanding treatment that violates a guideline, the treatment received is instrumented using measures of local practice style interacted with patient characteristics. The large number of possible instruments generated by this process is winnowed using the post-lasso two-stage least squares procedure suggested by Belloni et al. (2012). They find that patients who receive treatment that violates guidelines have higher health care costs, higher probabilities of self-harm, more ED visits, and more hospitalizations over the next two years. These results suggest that these patients would indeed be better off if doctors followed professional guidelines.

Abaluck et al. (2021) asks several additional questions about the use of guidelines. First, when guidelines change, how quickly do doctors update their practice style? Second, if doctors fail to update, is this because they are unaware of the changes or is it for other reasons? Third, are some violations of guidelines justified by treatment effect heterogeneity? They study the prescription of anticoagulants for patients with atrial fibrillation. Guidelines for treating these patients changed in 2006. Data from eight randomized controlled trials (RCTs) are available to try to explore treatment effect heterogeneity. They measure doctor awareness of the new procedures by using text mining of electronic medical records to find the first time the doctor mentioned them. After that date, the doctor is assumed to be aware of the new guidelines. The results suggest that doctors do move toward the new guidelines but that adherence is highly imperfect. They estimate that stricter adherence to the new guidelines could have prevented 24% more strokes. They also find that departures from the guidelines do not seem to be justified by heterogeneity in treatment effects. Shurtz et al. (2024) have a similar finding with respect to colonoscopies. They find that when a doctor's patient receives an unexpected colon cancer diagnosis, doctors are more likely to screen patients appropriately, but only for three months. Similarly, Singh (2021) shows that when obstetricians experience complications with one mode of delivery, they tend to switch to the other, but only temporarily. Even in cases where following guidelines has a clear health benefit, it appears to be difficult to achieve compliance.

Kowalski (2023) raises an additional issue—what if the guidelines are followed, but are flawed? She studies U.S. mammography screening guidelines, which specify that women between ages 40 and 50 can make an individual decision in consultation with their doctors about whether mammography is warranted. Other countries, including Canada, recommend against the screening of asymptomatic women aged 40 to 50. The data come from a large Canadian RCT. Women in the treatment group were offered mammograms between 40 and 50. The control group was not offered mammograms at those ages. A novel feature of her analysis is that she differentiates between the rates of overdiagnosis for women who always got a mammography regardless of their assignment to the treatment and control group; women who are more likely to get mammograms if they are in the treatment group (the compliers); and those who never received mammograms regardless of their treatment status (the noncompliers). She finds that under the voluntary screening regime, the women who are screened are disproportionately healthier and of higher socioeconomic status. Moreover, 14% of the cancers uncovered in the complier group are "overdiagnosed" in the sense that they were noninvasive cancers that would never have led to symptoms if they had remained undetected, while 36% of the cancers detected in the group that always got mammograms were overdiagnosed. She also discusses underdiagnosis but finds little evidence that cancers that would cause harm to the patient are being missed under the lighter screening regime. The results imply that, if compliers in the U.S. are similar to those in Canada, bringing the U.S. guidelines into compliance with those of other countries would be beneficial in the sense that it would eliminate overdiagnosis that leads to harmful overtreatment.

In sum, the limited economic research available suggests that guidelines have the potential to improve outcomes if doctors can be persuaded to follow them, and if they can be updated in a timely way when new knowledge becomes available. It is not known how current clinical practice is shaped by guidelines or what measures would be most effective in promoting adherence to guidelines. Finally, there has been little research on the socially optimal form of guidelines. Should they be very prescriptive (i.e., checklists), or should they be guardrails that discourage some treatments but allow flexibility in treatment choices within relatively broad limits? These are all important questions for future research.

5.3 Can technology improve medical decision making?

It may seem obvious that technology can improve medical decision making. For example, the invention of the mammogram meant that in many cases, doctors could tell whether a lump was likely to be cancerous or not. But as Kowalski's study illustrates, a new tool can be overused or underused. Moreover, the use of the tool may expose patients to other dangers, such as radiation, and unnecessary surgery or chemotherapy in the case of mammograms.[34] This Section focuses on technologies that have been touted as having the potential to revolutionize medicine including telemedicine (or telehealth), electronic medical records (EMRs), and prescription drug monitoring programs (PDMPs), and on the use of algorithms to assist decision making. Some of the many studies in these areas are summarized in Appendix Table 8.

Telehealth is a technology with potentially widespread effects on medical decision making. Zeltzer et al. (2023) evaluate the introduction of a device that facilitated telehealth primary care visits by allowing patients to collect and upload basic health data. The device reduced urgent care, ED, and inpatient visits and increased primary care visits, suggesting increases in the efficiency of medical care delivery. However, it also increased the use of antibiotics, which is concerning. Zeltzer et al. (2024) treat the COVID-19 pandemic as a shock that increased access to telemedicine in Israel in a long-lasting way. They find increases in primary care visits but a reduction in overall costs. There was no evidence of increases in missed diagnoses.

Dahlstrand (2022) suggests that telemedicine has the potential to improve patient health and reduce health disparities by allowing sick patients to access skilled doctors regardless of their location. She estimates that matching patients at risk for avoidable hospitalization with the most-skilled doctors would lead to an 8% reduction in such hospitalizations. However, it remains to be seen whether these kinds of hypothetical gains can be realized. Would less sick but privileged patients tolerate reduced access to the highest skilled physicians in order to accommodate the patients at highest risk?

Goetz (2023) examines the impact of a change in the algorithm that provides patients with information about online talk therapists. Initially, the platform only displayed providers in the patient's area. The change occurred in areas with fewer than 20 providers. It allowed patients in these areas to see information about providers in other areas. He shows that the change caused the most-skilled providers to stop offering sliding fees on-line, while less-skilled providers were more likely to exit the platform. Presumably, skilled therapists started receiving more requests for fee discounts, while less-skilled therapists lost patients to out-of-area providers. These results suggest that the market for telehealth is sensitive to seemingly small differences in platform architecture. Both Dahlstrand (2022) and Goetz (2023) also highlight the potential for telehealth to change the boundaries of health care markets. Such a change could affect provider competition and, potentially, patient health care utilization, costs, and health.

High-quality information about a patient's condition is essential to patient care, whether it is provided in person or via telemedicine. The development of EMRs may enable and incentivize doctors to keep better records and facilitate the coordination of care across providers. In some cases, EMRs are combined with other types of decision support tools. In the U.S., the use of EMRs was incentivized by the 2009 HITECH Act, which was itself part of the federal government's response to the Great Recession. The Act set goals for the adoption of EMRs and gave providers financial incentives to encourage them to meet these goals. In retrospect, it is unfortunate that the Act did not set standards for the interoperability of different EMR systems. Today, while most providers use EMRs, there are many incompatible programs in use, limiting the extent to which EMR adoption can reduce the fragmentation of care. Other countries, such as England, have also struggled to implement unified, interoperable systems (Wilson and Khansa, 2018).

Most economic studies of EMRs have focused on whether their adoption has improved the quality of care. Even in the absence of better care coordination, EMRs could improve the care provided by individual

[34]There is a large literature on the overuse of imaging technology more generally. For example, **?** compare bordering areas with and without certificate of need (CON) laws, which restrict the use of imaging technology. They find that CON laws reduce the probability of receiving low-value magnetic resonance imaging without affecting high-value imaging. However, the same laws reduce the probability of getting even high-value CT scans.

clinicians. By requiring doctors to fill in certain fields, an EMR might prompt them to think about attributes of patients or care options that they would otherwise have neglected. An EMR might also lead to better care coordination within a practice or hospital, which could improve outcomes. A third possibility is that a more comprehensive track record encourages doctors to take more care, lest they should be accused of malpractice. On the other hand, EMRs have proven unpopular with many clinicians, who complain of information overload. One survey of primary care physicians in the U.S. Veterans Health Administration found that 90% of doctors found the number of alerts that they received excessive. Over half of the respondents said that the flood of information increased the probability of overlooking important data (Singh et al., 2013).

In one of the first papers on this topic, McCullough et al. (2010) examined the impact of EMR adoption on hospital-level (and hospital reported) measures of the quality of care. They find that only two of the many measures they examined showed any impact. Agha (2014) uses individual-level Medicare claims data to examine the impact of EMR adoption in models with hospital fixed effects. She finds that adoption increased health care spending by 1.3%, but it had no impact on length of stay, intensity of care, care quality, re-admissions, or one-year mortality. In contrast to these two studies, Miller and Tucker (2011) use county-level data to examine the impact of EMR adoption over the 1995—2006 period. EMR adoption is instrumented using state medical privacy laws. They argue that by inhibiting the sharing of information, such laws make EMR adoption less attractive. They find that a 10% increase in EMR adoption reduces neonatal mortality by 3%. These reductions are due to prematurity and complications of labor and delivery and not to accidents, sudden infant death syndrome, or congenital defects. A caveat is that they cannot observe whether a particular baby was actually delivered in a hospital with EMRs, and there might be other changes in medical care in counties that happened to be rapid adopters of EMRs.

One interesting potential use of EMRs is to identify areas of concern so that they can be targeted for improvement. For example, in 2006, the state of California began an initiative to reduce maternal mortality. The first step was to identify hospitals with high rates and to determine the most important cause of maternal deaths in each hospital. This cause was then targeted. For example, if a lot of mothers were dying of hemorrhage, staff were trained to identify mothers at risk and a "crash cart" was assembled with everything necessary to treat maternal hemorrhage in one place (Main et al. (2020)). This initiative reduced maternal mortality in California by 65% from 2006 to 2016, while rates continued to increase in the rest of the U.S.[35]

Prescription drug monitoring programs (PDMPs) can be thought of as a specific and limited type of EMR. A PDMP is a state-level electronic registry of prescriptions for controlled drugs such as opioids and benzodiazepines. PDMPs can be searched by doctors, administrators, or law enforcement (depending on state rules) to identify patients or doctors who are using or prescribing drugs improperly. Because they are run at the state level, they come in many different flavors, but one of the most important distinctions is whether doctors are required to access the PDMP before prescribing. Several studies have found that the adoption of these “must access” PDMPs reduced prescribing of opioids but had limited impacts on outcomes such as overdose deaths (Buchmueller and Carey (2018); Sacks et al. (2021); Neumark and Savych (2023)). One possible reason that PDMP adoption might have limited initial effects on overdoses is that it may take some time for a new opioid prescription to lead to addiction and death, so that the standard difference-in-differences framework may not be well suited to capturing these delayed effects.

Alpert et al. (2024) interpret a must access PDMP as something that imposes an additional “hassle cost” on prescribing compared to a PDMP that is not use must access. They argue that if the PDMP operated

[35] See `https://www.cmqcc.org/who-we-are`.

mainly by providing information to prescribers about patients who were abusing opioids, then it should have no effect on opioid-naive patients, that is, patients who were not already taking opioids. However, they show that the adoption of a must access PDMP affects both opioid naive and non-opioid naive patients, though it affects the latter more. They also note that the patients who needed opioids the most, such as cancer patients, still received them, so increasing the cost of prescribing improved the targeting of treatment. They conclude that hassle costs, rather than increases in the information available to providers, explain most of the observed decline in opioid prescribing with must access PDMPs. Another interpretation of these results is that the mere implementation of a must access PDMP provides a signal to physicians about the risks associated with opioids.

In terms of other outcomes, Sacks et al. (2021) observe that PDMPs do not significantly affect "extreme use such as doctor shopping among new patients, because such behavior is very rare."[36] This finding is ironic because the idea that addicted patients were "doctor shopping" to obtain multiple prescriptions of dangerous medications was one of the prime motivations for the creation of PDMPs.

Another technological approach to improving decision making is to use an algorithmic decision tool. Interest in using algorithms to assist physician decision making dates back at least to Meehl's 1954 book on the subject and the seminal article by Ledley and Lusted (1959) in *Science*. It is worthwhile to briefly discuss what an algorithm is, especially given the recent interest in large language models and their potential impact on labor markets.

All algorithms are functions that take in numerical data and produce a numerical output. For example, in the case of large language models, the text is mapped into a high dimensional vector space ($\Re^n$,where n is a large number) and then transformed via a sequence of mathematical operations. In the context of our model, the output could be the probability that the intensive treatment is best, $\rho(\vec{x}_i)$, where x_i is a vector representing all the information known about patient i. An algorithm will recommend intensive treatment if and only if the probability of intensive treatment is greater than one-half ($\rho(\vec{x}_i) > 1/2$).[37]

Humans also make decisions based on data. Moreover, humans can quickly process vast quantities of information through the visual field. Decades of research has shown that, in contrast to computers, humans cannot rapidly process large volumes of *numerical* information. When the numerical information provides a more accurate assessment of the benefits from a decision, then algorithms, even algorithms based on simple linear regressions, can perform better than a human decision maker.[38]

When good numerical data is available, we should expect algorithms to provide high quality recommendations that can improve on human decision makers. Ludwig et al. (2024) point to the algorithm Mullainathan and Obermeyer (2022) developed to predict who should be tested for heart attacks and argue that the adoption of such an algorithm would amount to a "free lunch" in the sense that the social benefit would greatly outweigh the cost.

Yet, since humans are capable of processing large volumes of visual data and making decisions in real time, a good doctor can tell at a glance that a wound is infected or that a patient has hepatitis. The fact that humans are very good at processing visual information implies that in some cases the doctor is simply the most efficient agent to collect and act on information. For example, a patient coming into an ED may immediately require intravenous fluids. Getting the person's weight and vital signs for the EMR takes time that might not be available. The attending doctor can estimate the patient's weight and condition in less

[36]Sacks et al. (2021), page 10297.

[37]See Devroye et al. (1996) on the mathematics of machine learning. Bengio et al. (2021) provide an up-to-date discussion of machine learning by three seminal contributors to the field.

[38]Kahneman (2003) noted in his Noble Prize lecture that he first recognized this point in the 1950s while working for the Israeli military. The seminal contribution by Dawes et al. (1989) makes this point in the context of medical decision making.

than a second, and then order or execute treatment. As Kahneman and Klein (2009) observe, there are many examples of experts with extraordinarily high levels of skill, and hence, both algorithms and skilled experts can play a role in improving decision making. At the same time, as the evidence reviewed above illustrates, there is a great deal of variation in doctor skill. The question then is how best to incorporate the benefits of well-designed algorithms while also exploiting the knowledge of highly skilled doctors.

This problem turns out to be quite difficult. Agarwal et al. (2023) conducted a randomized experiment with radiologists who were asked to retrospectively diagnose patients in a laboratory setting that resembled their usual working environment. In some cases, they received only an x-ray, while in other cases, they were given either an AI prediction, additional contextual information about the patient's history that was not considered by the AI tool, or both. The AI algorithm used has been shown to perform similarly to professional radiologists. The experimental subjects' diagnoses were then compared to "ground truth" derived using the opinions of five expert radiologists. Agarwal et al. (2023) find that giving radiologists the AI prediction did not improve diagnostic accuracy, while giving them additional contextual information did. They estimate a model of belief updating that suggests that clinicians erroneously treat the AI prediction as independent of their own information, which causes it to bias their decision making. They argue that better results could have been achieved by using the AI prediction in cases in which the tool had high confidence and allowing humans to make decisions without AI assistance in all other cases.

The problem of how to effectively combine algorithmic information and expert opinion arises in many other settings. For instance, Stevenson and Doleac (2022) find that judges given algorithmic assessments of the probability of recidivism change their sentencing decisions, but that use of the tool did not either reduce incarceration or improve public safety. Judges deviated from the algorithm in a way that increased incarceration but also reduced recidivism. Hoffman et al. (2018) look at manager hiring decisions before and after the introduction of formal job testing algorithms. They find that managers who overrule the algorithmic recommendation hire worse people on average. Rambachan (2024) adds to the literature on bail decisions, arguing that well-designed algorithms can improve judicial decisions.

The performance of AI models currently in clinical use is similarly mixed. Obermeyer et al. (2019) describe an algorithm that identified at-risk patients by calculating expected total medical expenditures. Because more is spent on white patients conditional on their underlying health conditions, such an algorithm will tend to short-change Black patients. One way to think about the problem is that the algorithm was trained on medical expenditure data that is biased in favor of white patients. The authors also note that it may be easier to correct such a problem in an algorithm than it is to get human decision makers to show less bias in the allocation of treatments.

Manz et al. (2023) conducted a large randomized trial to see whether a machine-learning generated nudge could encourage clinicians to engage in end-of-life conversations with terminally ill cancer patients. They find an increase in such conversations and a reduction in systemic cancer therapy at the end of life, but no change in hospice, length of stay, or intensive-care admission at the end of life.

Using data from one of the largest purveyors of EMRs, EPIC, Wong et al. (2021) find that an AI tool for diagnosing sepsis that is used in hundreds of hospitals performed poorly in a large teaching hospital setting. It failed to identify 67% of patients with sepsis even though it generated an alert for 18% of all patients. Lyons et al. (2023) followed up on this finding by examining the performance of the tool in nine networked hospitals. They find that the tool did better in hospitals treating patients who are less sick and have a lower average probability of sepsis.

As this example illustrates, even if an algorithm is trained on big data, it may not perform very well if the sample at hand is different from the one used to train the algorithm. Although economists have been aware of the selection problem since the famous work of Roy (1951) on wages and the self-selection of workers to occupations, awareness of the selection problem in the machine learning literature is very recent (see Athey and Imbens (2019)). Many modern machine learning algorithms in medicine have access to large amounts of data, with patients who are allocated to different treatments. The problem is that if one does not incorporate the allocation (selection) mechanism in the machine learning model, then the predicted effects of treatment may be incorrect. For example, if clinicians only give an experimental treatment to the patients they believe are most likely to recover, then the effectiveness of the treatment is likely to be overstated.

Moreover, Rambachan and Roth (2020) show that even if one knows the direction of the selection bias in the underlying data, the bias in the algorithm can be in any direction. This observation highlights the point that learning from large datasets requires more than simply choosing the right algorithm. It also entails understanding how the sample is selected and testing that the results apply in different settings. In the real world, an algorithm is trained and deployed in one setting, and then others may try to deploy it in a new setting where variables are coded differently, data are missing, or the initial investigators are no longer involved. It is little wonder that the algorithm may not perform well in these circumstances.

In summary, these three new technologies, telemedicine, EMRs, and algorithmic decision tools, have considerable promise, but the available evidence suggests that the details of how they are implemented really matter. More research is required to understand how to use them to actually improve patient welfare.

6 Conclusions and Suggestions for Future Research

In a world where there was little that could be done for most ailments, there were few consequential decisions to be made. Today, medical decision making matters more than ever. The model of medical decision making that we have outlined has several moving parts. Doctors are assumed to care about patient welfare but also about their own welfare, which makes them imperfect agents. Doctors arrive at the bedside with a given training and experience, which results in a set of skills as well as prior beliefs about proper treatment. As humans, doctors are influenced by fatigue, time pressures, emotional states, prejudices, and peer effects. They may rely on simple decision rules in cases where more focused attention would improve outcomes. At present, no one has estimated a model that parses out the roles of doctor diagnostic skill (γ_j), the impact of procedural skill as it effects the relative effectiveness of nonintensive and intensive treatments ($\Delta_{LNIj}/\Delta_{HIj}$), pecuniary factors ($\delta_{tj}$), differences in patient populations (α_i), doctor beliefs about patient populations (p_{Lj}/p_{Hj}), and the resulting decision thresholds that doctors (τ_j) set. As we have highlighted, in order to be tractable, existing models shut down one or more of these channels. Hence, estimating a richer model is a potentially useful direction for future research.

The fact that there are so many factors that affect medical decision making suggests that there is no one policy lever that will optimize care. In particular, the research reviewed here indicates that it can be difficult to tweak payment systems in a way that will have unambiguously positive effects on the allocation of medical care. Future work on the impacts of changes in payment systems (and other levers) should pay careful attention to their welfare consequences and incorporate heterogeneity in the effects on patients.

Other important areas for future work include research on the effectiveness of medical training that actually pays attention to the content of training at the undergraduate level, medical school, residency, or in continuing education. Existing studies tend to focus on crude measures such as years of training or type/rank of medical school.

Chronic doctor shortages in many countries suggest that there will be continuing demand for the services of even the least skilled physicians, which may attenuate any incentives for continuous skill improvement. Reforms that reduce physician burnout and exit would increase the supply of physicians. In turn, a larger physician supply might allow for further reductions in time pressures and burnout, but there has been little research on this question. It would also be interesting to see research on recent efforts to diversify the medical workforce by encouraging minority applicants to attend medical schools by waiving tuition or by subsidizing doctors in underserved areas.

Short anti-bias trainings offer an interesting case in which the impact of a specific form of training has been evaluated and found to have little impact on physician behavior. Vela et al. (2022)'s hypothesis that the effect of anti-bias training is counteracted by the messages implicit in the rest of a doctor's training suggests that it is necessary to better understand doctor training as a whole. Enhancing medical decision making by improving the concordance between the characteristics of doctors and patients will also take a long time. Research into other ways to enhance sympathy and communication between doctors and patients is sorely needed.

The fact that poor medical decision making is difficult to address with payment reforms or training (given what we now know about training effects) accounts for much of the excitement about guidelines, algorithms, and other emerging health care technologies among health economists. As researchers, we tend to have faith in the efficacy of providing information to economic agents, but the evidence reviewed here indicates that doctors pay more attention to some types of new information than others. Information provision alone does not eliminate undesirable variations in practice and does not always even lead to changes in the right direction. Key questions going forward are what factors make information salient and how these factors interact with physician characteristics.

Research suggests that adherence to clinical guidelines is helpful for patients, at least where the guidelines themselves represent best practice. But it is not known how current clinical practice is shaped by guidelines or what measures are most effective in promoting adherence to guidelines. There has also been little economic research on designing effective guidelines. Should they be very prescriptive (e.g. checklists), or should they be more in the nature of guardrails that forbid some treatments but allow flexibility within relatively broad limits? Are optimal guidelines different for simple versus complex cases?

Telemedicine, EMRs, and algorithmic decision tools have considerable promise, but we do not yet understand how to implement them to assist optimal decision making. Like older medical technologies, these new tools can be overused, underused, and can lead to harmful consequences for patients when used inappropriately. Understanding how humans can interact with the tools to produce better outcomes is a first order question. In the real world, a tool that worked very well in the setting it was designed for may be difficult to implement and produce substandard decisions in a different setting. Designing algorithms that are easy to customize and implement across settings, and which take into account the way that humans interact with machines, is an important priority for future work.

Health care data offer unique opportunities to observe both physician decisions and their consequences for patients. The literature we discuss speaks to questions about labor productivity, organizational economics, and the use of technology that are often difficult to analyze in other settings, if only because it is usually so difficult to see the downstream consequences of an expert decision. Many of the themes we highlight here may be relevant to other labor markets with high-skilled workers. Hence, it is interesting to ask which insights about factors that affect medical decision making can be transferred to other settings with highly skilled decision makers.

The empirical work we have reviewed wrestles with ubiquitous selection problems. Patients select doctors and may also choose procedures. Doctors may select patients. Medical schools and training programs select applicants. Doctors select peers. The most successful papers in this literature identify situations that approximate random assignment to doctors, treatments, or to a particular medical team in order to achieve causal identification.[39] This work has shown both that different doctors treat medically similar patients differently, and that individual doctors often treat similar patients differently depending on patient characteristics such as age, race, and gender, or depending on time-varying doctor-specific factors such as the time left in their shift or the presence of peers. One caveat is that much of this work focuses on elderly Medicare patients for reasons of data availability, so extending these results to other populations and settings would be useful. A second caveat is that even when we can identify causal effects, it is difficult to understand the precise mechanisms and motivations underlying doctor decisions. Better understanding of mechanisms is necessary for the development of effective interventions to improve doctor decision making.

References

Abaluck, J., L. Agha, J. Chan, David C., D. Singer, and D. Zhu (2021). Fixing Misallocation with Guidelines: Awareness vs. Adherence. Working Paper 27467, National Bureau of Economic Research.

Abaluck, J., L. Agha, C. Kabrhel, A. Raja, and A. Venkatesh (2016, December). The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care. *American Economic Review 106*(12), 3730–3764.

Agarwal, N., A. Moehring, P. Rajpurkar, and T. Salz (2023, July). Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. Working Paper 31422, National Bureau of Economic Research.

Agha, L. (2014, March). The effects of health information technology on the costs and quality of medical care. *Journal of Health Economics 34*, 19–30.

Agha, L. and D. Molitor (2018, March). The Local Influence Of Pioneer Investigators On Technology Adoption: Evidence From New Cancer Drugs. *The Review of Economics and Statistics 100*(1), 29–44.

Agha, L. and D. Zeltzer (2022, May). Drug Diffusion through Peer Networks: The Influence of Industry Payments. *American Economic Journal: Economic Policy 14*(2), 1–33.

Ahammer, A. and T. Schober (2020). Exploring variations in health-care expenditures—What is the role of practice styles? *Health Economics 29*(6), 683–699.

Ahomäki, I., V. Pitkänen, A. Soppi, and L. Saastamoinen (2020, July). Impact of a physician-targeted letter on opioid prescribing. *Journal of Health Economics 72*, 1–22.

Alexander, D. (2020, November). How Do Doctors Respond to Incentives? Unintended Consequences of Paying Doctors to Reduce Costs. *Journal of Political Economy 128*(11), 4046–4096.

Alexander, D. and J. Currie (2017, May). Are publicly insured children less likely to be admitted to hospital than the privately insured (and does it matter)? *Economics & Human Biology 25*, 33–51.

Alexander, D. and M. Schnell (2024). The Impacts of Physician Payments on Patient Access, Use, and Health. *American Economic Journal: Applied Economics*.

Almond, D., J. J. Doyle, Jr., A. E. Kowalski, and H. Williams (2010, May). Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns. *The Quarterly Journal of Economics 125*(2), 591–634.

[39]See Holland (1986) for discussion of the basic concepts, and Imbens and Rubin (2015) for a book length treatment of causal identification.

Almond, D., J. J. Doyle, Jr., A. E. Kowalski, and H. Williams (2011, November). The Role of Hospital Heterogeneity in Measuring Marginal Returns to Medical Care: A Reply to Barreca, Guldi, Lindo, and Waddell. *The Quarterly Journal of Economics 126*(4), 2125–2131.

Alpert, A., S. Dykstra, and M. Jacobson (2024, February). Hassle Costs versus Information: How Do Prescription Drug Monitoring Programs Reduce Opioid Prescribing? *American Economic Journal: Economic Policy 16*(1), 87–123.

Alsan, M., O. Garrick, and G. Graziani (2019, December). Does Diversity Matter for Health? Experimental Evidence from Oakland. *American Economic Review 109*(12), 4071–4111.

Alsan, M. and M. Wanamaker (2018, February). Tuskegee and the Health of Black Men. *The Quarterly Journal of Economics 133*(1), 407–455.

Angerer, S., C. Waibel, and H. Stummer (2019, October). Discrimination in Health Care: A Field Experiment on the Impact of Patients' Socioeconomic Status on Access to Care. *American Journal of Health Economics 5*(4), 407–427.

Arnold, D., W. Dobbie, and P. Hull (2022, September). Measuring Racial Discrimination in Bail Decisions. *American Economic Review 112*(9), 2992–3038.

Athey, S. and G. W. Imbens (2019, August). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics 11*(1), 685–725.

Avdic, D., S. von Hinke, B. Lagerqvist, C. Propper, and J. Vikström (2024, March). Do responses to news matter? Evidence from interventional cardiology. *Journal of Health Economics 94*, 102846.

Azoulay, P. (2002). Do Pharmaceutical Sales Respond to Scientific Evidence? *Journal of Economics & Management Strategy 11*(4), 551–594.

Badinski, I., A. Finkelstein, M. Gentzkow, and P. Hull (2023, October). Geographic Variation in Healthcare Utilization: The Role of Physicians. Working Paper 31749, National Bureau of Economic Research.

Balogh, E. P., B. T. Miller, and J. R. Ball (Eds.) (2015, December). *Improving Diagnosis in Health Care.* Washington, D.C.: National Academies Press.

Barreca, A. I., M. Guldi, J. M. Lindo, and G. R. Waddell (2011, November). Saving Babies? Revisiting the effect of very low birth weight classification. *The Quarterly Journal of Economics 126*(4), 2117–2123.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica 80*(6), 2369–2429.

Bengio, Y., Y. Lecun, and G. Hinton (2021, July). Deep learning for AI. *Communications of the ACM 64*(7), 58–65.

Berndt, E. R., R. S. Gibbons, A. Kolotilin, and A. L. Taub (2015, March). The heterogeneity of concentrated prescribing behavior: Theory and evidence from antipsychotics. *Journal of Health Economics 40*, 26–39.

Betz, T., S. J. Cook, and F. M. Hollenbach (2018, October). On the Use and Abuse of Spatial Instruments. *Political Analysis 26*(4), 474–479.

Bisgaier, J. and K. V. Rhodes (2011, June). Auditing Access to Specialty Care for Children with Public Insurance. *New England Journal of Medicine 364*(24), 2324–2333.

Brekke, K. R., T. H. Holmås, K. Monstad, and O. R. Straume (2018). Socio-economic status and physicians' treatment decisions. *Health Economics 27*(3), e77–e89.

Brekke, K. R., T. H. Holmås, K. Monstad, and O. R. Straume (2019, July). Competition and physician behaviour: Does the competitive environment affect the propensity to issue sickness certificates? *Journal of Health Economics 66*, 117–135.

Buchmueller, T. C. and C. Carey (2018, February). The Effect of Prescription Drug Monitoring Programs on Opioid Utilization in Medicare. *American Economic Journal: Economic Policy 10*(1), 77–112.

Button, P., E. Dils, B. Harrell, L. Fumarco, and D. Schwegman (2020, December). Gender Identity, Race, and Ethnicity Discrimination in Access to Mental Health Care: Preliminary Evidence from a Multi-Wave Audit Field Experiment. Working Paper 28164, National Bureau of Economic Research.

Cabral, M., C. Carey, and S. Miller (2021, November). The Impact of Provider Payments on Health Care Utilization of Low-Income Individuals: Evidence from Medicare and Medicaid.

Cabral, M. and M. Dillender (2024, February). Gender Differences in Medical Evaluations: Evidence from Randomly Assigned Doctors. *American Economic Review 114*(2), 462–499.

Carey, C., M. Daly, and J. Li (2024, April). Nothing for Something: Marketing Cancer Drugs to Physicians Increases Prescribing Without Improving Mortality.

Carey, C., E. M. J. Lieber, and S. Miller (2021, May). Drug firms' payments and physicians' prescribing behavior in Medicare Part D. *Journal of Public Economics 197*, 104402.

Chan, D. C. (2016, June). Teamwork and Moral Hazard: Evidence from the Emergency Department. *Journal of Political Economy 124*(3), 734–770.

Chan, D. C. (2018). The efficiency of slacking off: Evidence from the emergency department. *Econometrica 86*(3), 997–1030.

Chan, D. C. (2021, February). Influence and Information in Team Decisions: Evidence from Medical Residency. *American Economic Journal: Economic Policy 13*(1), 106–137.

Chan, D. C., M. Gentzkow, and C. Yu (2022, April). Selection with Variation in Diagnostic Skill: Evidence from Radiologists. *The Quarterly Journal of Economics 137*(2), 729–783.

Chan, David C., J. and Y. Chen (2022, October). The Productivity of Professions: Evidence from the Emergency Department. Working Paper 30608, National Bureau of Economic Research.

Chandra, A. and J. Skinner (2012, July). Technology Growth and Expenditure Growth in Health Care. *Journal of Economic Literature 50*(3), 645–680.

Chandra, A. and D. O. Staiger (2007). Productivity spillovers in health care: Evidence from the treatment of heart attacks. *Journal of Political Economy 115*(1), pp.103–140.

Chandra, A. and D. O. Staiger (2010, September). Identifying Provider Prejudice in Healthcare. Working Paper 16382, National Bureau of Economic Research.

Chandra, A. and D. O. Staiger (2020, May). Identifying Sources of Inefficiency in Healthcare. *The Quarterly Journal of Economics 135*(2), 785–843.

Chen, A. and D. N. Lakdawalla (2019, March). Healing the poor: The influence of patient socioeconomic status on physician supply responses. *Journal of Health Economics 64*, 43–54.

Chen, Y. (2021, December). Team-Specific Human Capital and Team Performance: Evidence from Doctors. *American Economic Review 111*(12), 3923–3962.

Chernew, M., Z. Cooper, E. L. Hallock, and F. Scott Morton (2021, March). Physician agency, consumerism, and the consumption of lower-limb MRI scans. *Journal of Health Economics 76*, 102427.

Chodick, G., Y. Goldstein, I. Shurtz, and D. Zeltzer (2023, March). Challenging Encounters and Within-Physician Practice Variability. *The Review of Economics and Statistics*, 1–27.

Chorniy, A., J. Currie, and L. Sonchak (2018, July). Exploding asthma and ADHD caseloads: The role of medicaid managed care. *Journal of Health Economics 60*, 1–15.

Chowdhury, M. M., H. Dagash, and A. Pierro (2007). A systematic review of the impact of volume of surgery and specialization on patient outcome. *BJS (British Journal of Surgery) 94*(2), 145–161.

Chu, B., B. Handel, J. Kolstad, J. Knecht, U. Malmendier, and F. Matejka (2024). Cognitive Capacity, Fatigue and Decision Making: Evidence from the Practice of Medicine. Technical report, UC Berkeley.

Chyn, E., B. Frandsen, and E. C. Leslie (2024, April). Examiner and Judge Designs in Economics: A Practitioner's Guide.

Clemens, J. and J. D. Gottlieb (2014, April). Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health? *American Economic Review 104*(4), 1320–1349.

Clemens, J., P.-T. Léger, Y. Nandan, and R. Town (2024, October). Physician Practice Preferences and Healthcare Expenditures: Evidence from Commercial Payers.

Costa-Ramón, A. M., A. Rodríguez-González, M. Serra-Burriel, and C. Campillo-Artero (2018, May). It's about time: Cesarean sections and neonatal health. *Journal of Health Economics 59*, 46–59.

Coudin, E., A. Pla, and A.-L. Samson (2015). GP responses to price regulation: Evidence from a French nationwide reform. *Health Economics 24*(9), 1118–1130.

Coussens, S. (2018, August). Behaving Discretely: Heuristic Thinking in the Emergency Department.

Cuddy, E. and J. Currie (2020, September). Treatment of mental illness in American adolescents varies widely within and across areas. *Proceedings of the National Academy of Sciences 117*(39), 24039–24046.

Cuddy, E. and J. Currie (2024). Rules vs. Discretion: Treatment of Mental Illness in U.S. Adolescents. *Journal of Political Economy*.

Currie, J., A. Li, and M. Schnell (2023, January). The Effects of Competition on Physician Prescribing. Working Paper 30889, National Bureau of Economic Research.

Currie, J., W. B. MacLeod, and M. Ouyang (2024, May). Spillovers and Training Effects on Mental Health Prescribing for Children. *AEA Papers and Proceedings 114*, 394–400.

Currie, J., W. B. MacLeod, and J. Van Parys (2016, May). Provider practice style and patient health outcomes: The case of heart attacks. *Journal of Health Economics 47*, 64–80.

Currie, J. and H. Schwandt (2021, May). The Opioid Epidemic Was Not Caused by Economic Distress but by Factors That Could Be More Rapidly Addressed. *The ANNALS of the American Academy of Political and Social Science 695*(1), 276–291.

Currie, J. and J. Zhang (2023, February). Doing More with Less: Predicting Primary Care Provider Effectiveness. *The Review of Economics and Statistics*, 1–45.

Currie, J. M. and W. B. MacLeod (2008, May). First do no harm? Tort reform and birth outcomes. *Quarterly Journal of Economics 123*(2), 795–830.

Currie, J. M. and W. B. MacLeod (2017, January). Diagnosis and unnecessary procedure use: Evidence from C-section. *Journal of Labor Economics 35*(1), 1–42.

Currie, J. M. and W. B. MacLeod (2020, May). Understanding doctor decision making: The case of depression treatment. *Econometrica 88*(3), 847–878.

Cutler, D., J. S. Skinner, A. D. Stern, and D. Wennberg (2019, February). Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending. *American Economic Journal: Economic Policy 11*(1), 192–221.

Cutler, D. M. (2014). *The Quality Cure: How Focusing on Health Care Quality Can Save Your Life and Lower Spending Too*, Volume 9 of *The Aaron Wildavsky Forum for Public Policy.* Berkeley: University of California Press.

Dahlstrand, A. (2022). Defying distance? The provision of services in the digital age. Technical Report CEPDP1889, Centre for Economic Performance, LSE.

Dawes, R. M., D. Faust, and P. E. Meehl (1989, March). Clinical Versus Actuarial Judgment. *Science 243*(4899), 1668–1674.

DeCicca, P., M. Isabelle, and N. Malak (2024). How do physicians respond to new medical research? *Health Economics 33*(10), 2206–2228.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition.* New York, NY: Springer-Verlag.

Ding, Y. and C. Liu (2021, December). Alternative payment models and physician treatment decisions: Evidence from lower back pain. *Journal of Health Economics 80.*

Doctor, J. N., A. Nguyen, R. Lev, J. Lucas, T. Knight, H. Zhao, and M. Menchine (2018, August). Opioid prescribing decreases after learning of a patient's fatal overdose. *Science 361*(6402), 588–590.

Doyle, J. J., S. M. Ewer, and T. H. Wagner (2010, December). Returns to physician human capital: Evidence from patients randomized to physician teams. *Journal of Health Economics 29*(6), 866–882.

Doyle, Joseph J., Jr. (2020, July). Physician Characteristics and Patient Survival: Evidence from Physician Availability. Working Paper 27458, National Bureau of Economic Research.

Dubois, P. and T. Tunçel (2021, July). Identifying the effects of scientific information and recommendations on physicians' prescribing behavior. *Journal of Health Economics 78*, 102461.

Dunn, A., J. D. Gottlieb, A. H. Shapiro, D. J. Sonnenstuhl, and P. Tebaldi (2024, February). A Denial a Day Keeps the Doctor Away. *The Quarterly Journal of Economics 139*(1), 187–233.

Eli, S., T. D. Logan, and B. Miloucheva (2019, May). Physician Bias and Racial Disparities in Health: Evidence from Veterans' Pensions. Working Paper 25846, National Bureau of Economic Research.

Epstein, A. J., S. Nicholson, and D. A. Asch (2016, February). The Production of and Market for New Physicians' Skill. *American Journal of Health Economics 2*(1), 41–65.

Facchini, G. (2022). Forgetting-by-not-doing: The case of surgeons and Cesarean sections. *Health Economics 31*(3), 481–495.

Fadlon, I. and J. van Parys (2020, May). Primary care physician practice styles and patient care: Evidence from physician exits in Medicare. *Journal of Health Economics 71*, 102304.

Fawcett, T. (2006). An introduction of ROC analysis. *Pattern Recognition Letters 27*, 861–874.

Feng, K., H. Hong, K. Tang, and J. Wang (2023, June). Statistical Tests for Replacing Human Decision Makers with Algorithms.

Frakes, M. (2013, February). The Impact of Medical Liability Standards on Regional Variations in Physician Behavior: Evidence from the Adoption of National-Standard Rules. *American Economic Review 103*(1), 257–276.

Frakes, M. D. and J. Gruber (2022, December). Racial Concordance and the Quality of Medical Care: Evidence from the Military. Working Paper 30767, National Bureau of Economic Research.

Freedman, S., E. Golberstein, T.-Y. Huang, D. J. Satin, and L. B. Smith (2021, May). Docs with their eyes on the clock? The effect of time pressures on primary care productivity. *Journal of Health Economics 77*, 102442.

Geiger, C. K., M. A. Clapp, and J. L. Cohen (2021, December). Association of Prenatal Care Services, Maternal Morbidity, and Perinatal Mortality With the Advanced Maternal Age Cutoff of 35 Years. *JAMA Health Forum 2*(12), e214044.

Gibbons, R. D., C. H. Brown, K. Hur, S. M. Marcus, D. K. Bhaumik, J. A. Erkens, R. M. Herings, and J. J. Mann (2007, September). Early Evidence on the Effects of Regulators' Suicidality Warnings on SSRI Prescriptions and Suicide in Children and Adolescents. *American Journal of Psychiatry 164*(9), 1356–1363.

Goetz, D. (2023, July). Telemedicine competition, pricing, and technology adoption: Evidence from talk therapists. *International Journal of Industrial Organization 89*, 102956.

Gowrisankaran, G., K. Joiner, and P. T. Léger (2022, October). Physician Practice Style and Healthcare Costs: Evidence from Emergency Departments. *Management Science*.

Goyal, M. K., N. Kuppermann, S. D. Cleary, S. J. Teach, and J. M. Chamberlain (2015, November). Racial Disparities in Pain Management of Children With Appendicitis in Emergency Departments. *JAMA Pediatrics 169*(11), 996–1002.

Greenwood, B. N., S. Carnahan, and L. Huang (2018, August). Patient–physician gender concordance and increased mortality among female heart attack patients. *Proceedings of the National Academy of Sciences 115*(34), 8569–8574.

Greenwood, B. N., R. R. Hardeman, L. Huang, and A. Sojourner (2020, September). Physician–patient racial concordance and disparities in birthing mortality for newborns. *Proceedings of the National Academy of Sciences 117*(35), 21194–21200.

Gruber, J., T. P. Hoe, and G. Stoye (2021, May). Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers. *The Review of Economics and Statistics*, 1–45.

Gupta, A. (2021, April). Impacts of Performance Pay for Hospitals: The Readmissions Reduction Program. *American Economic Review 111*(4), 1241–1283.

Hammad, T. A., T. Laughren, and J. Racoosin (2006, March). Suicidality in pediatric patients treated with antidepressant drugs. *Archives of General Psychiatry 63*(3), 332–339.

Handel, B. and K. Ho (2021, January). The industrial organization of health care markets. In K. Ho, A. Hortaçsu, and A. Lizzeri (Eds.), *Handbook of Industrial Organization*, Volume 5 of *Handbook of Industrial Organization, Volume 5*, pp. 521–614. Elsevier.

Hill, A. J., D. B. Jones, and L. Woodworth (2023, December). Physician-patient race-match reduces patient mortality. *Journal of Health Economics 92*.

Hoffman, K. M., S. Trawalter, J. R. Axt, and M. N. Oliver (2016, April). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences 113*(16), 4296–4301.

Hoffman, M., L. B. Kahn, and D. Li (2018, May). Discretion in Hiring. *The Quarterly Journal of Economics 133*(2), 765–800.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Howard, D. H. and J. Hockenberry (2019, June). Physician age and the abandonment of episiotomy. *Health Services Research 54*(3), 650–657.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences.* Cambridge University Press.

Johnson, E. M. and M. M. Rehavi (2016, February). Physicians Treating Physicians: Information and Incentives in Childbirth. *American Economic Journal: Economic Policy 8*(1), 115–141.

Kahneman, D. (2003). Maps of bounded rationality: A perspective on intuitive judgment and choice. In R. Fangsmyr (Ed.), *The Nobel Prizes 2002*, Chapter Autobiography, pp. 449–489. Stockholm, Sweden: Almqvist & Wiksell International.

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

Kahneman, D., P. Slovic, and A. Tversky (1982, April). *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge University Press.

Kessler, D. and M. McClellan (1996). Do doctors practice defensive medicine? *Quarterly Journal of Economics 111*(2), 353–90.

Kolstad, J. T. (2013, December). Information and Quality When Motivation Is Intrinsic: Evidence from Surgeon Report Cards. *American Economic Review 103*(7), 2875–2910.

Kowalski, A. E. (2023, January). Behaviour within a Clinical Trial and Implications for Mammography Guidelines. *The Review of Economic Studies 90*(1), 432–462.

Ledley, R. S. and L. B. Lusted (1959, July). Reasoning Foundations of Medical Diagnosis. *Science 130*(3366), 9–21.

Ludwig, J., S. Mullainathan, and A. Rambachan (2024, February). The Unreasonable Effectiveness of Algorithms.

Ly, D. P., P. G. Shekelle, and Z. Song (2023, August). Evidence for Anchoring Bias During Physician Decision-Making. *JAMA Internal Medicine 183*(8), 818–823.

Lyons, P. G., M. R. Hofford, S. C. Yu, A. P. Michelson, P. R. O. Payne, C. L. Hough, and K. Singh (2023, June). Factors Associated With Variability in the Performance of a Proprietary Sepsis Prediction Model Across 9 Networked Hospitals in the US. *JAMA Internal Medicine 183*(6), 611–612.

Main, E. K., S.-C. Chang, R. Dhurjati, V. Cape, J. Profit, and J. B. Gould (2020, July). Reduction in racial disparities in severe maternal morbidity from hemorrhage in a large-scale quality improvement collaborative. *American journal of obstetrics and gynecology 223*(1), 123.e1–123.e14.

Manz, C. R., Y. Zhang, K. Chen, Q. Long, D. S. Small, C. N. Evans, C. Chivers, S. H. Regli, C. W. Hanson, J. E. Bekelman, J. Braun, C. A. L. Rareshide, N. O'Connor, P. Kumar, L. M. Schuchter, L. N. Shulman, M. S. Patel, and R. B. Parikh (2023, March). Long-term Effect of Machine Learning–Triggered Behavioral Nudges on Serious Illness Conversations and End-of-Life Outcomes Among Patients With Cancer: A Randomized Clinical Trial. *JAMA Oncology 9*(3), 414–418.

Marquardt, K. (2021). Physician Practice Style for Mental Health Conditions: The Case of ADHD. Technical report.

McCullough, J. S., M. Casey, I. Moscovice, and S. Prasad (2010, April). The Effect Of Health Information Technology On Quality In U.S. Hospitals. *Health Affairs 29*(4), 647–654.

McDevitt, R. C. and J. W. Roberts (2014). Market structure and gender disparity in health care: Preferences, competition, and quality of care. *The RAND Journal of Economics 45*(1), 116–139.

McKibbin, R. (2023, July). The effect of RCTs on drug demand: Evidence from off-label cancer drugs. *Journal of Health Economics 90*, 102779.

Meehl, P. E. (1954). *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence. Minneapolis, MN, US: University of Minnesota Press.

Mello, M. M., M. D. Frakes, E. Blumenkranz, and D. M. Studdert (2020, January). Malpractice Liability and Health Care Quality: A Review. *JAMA 323*(4), 352–366.

Miller, A. R. and C. E. Tucker (2011, April). Can Health Care Information Technology Save Babies? *Journal of Political Economy 119*(2), 289–324.

Molitor, D. (2018, February). The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration. *American Economic Journal: Economic Policy 10*(1), 326–356.

Mullainathan, S. and Z. Obermeyer (2022, April). Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. *The Quarterly Journal of Economics 137*(2), 679–727.

Neumark, D. and B. Savych (2023, June). Effects of Opioid-Related Policies on Opioid Utilization, Nature of Medical Care, and Duration of Disability. *American Journal of Health Economics 9*(3), 331–373.

Newham, M. and M. Valente (2024, May). The cost of influence: How gifts to physicians shape prescriptions and drug costs. *Journal of Health Economics 95*, 102887.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan (2019, October). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447–453.

Olenski, A. R., A. Zimerman, S. Coussens, and A. B. Jena (2020, February). Behavioral Heuristics in Coronary-Artery Bypass Graft Surgery. *New England Journal of Medicine 382*(8), 778–779.

Parapia, L. A. (2008). History of bloodletting by phlebotomy. *British Journal of Haematology 143*(4), 490–495.

Persson, E., K. Barrafrem, A. Meunier, and G. Tinghög (2019). The effect of decision fatigue on surgeons' clinical decision making. *Health Economics 28*(10), 1194–1203.

Persson, P., X. Qiu, and M. Rossin-Slater (2021, January). Family Spillover Effects of Marginal Diagnoses: The Case of ADHD. Working Paper 28334, National Bureau of Economic Research.

Rambachan, A. (2024, May). Identifying Prediction Mistakes in Observational Data. *The Quarterly Journal of Economics*, qjae013.

Rambachan, A. and J. Roth (2020, December). Bias In, Bias Out? Evaluating the Folk Wisdom.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), pp.135–146.

Sabin, J. A. and A. G. Greenwald (2012, May). The Influence of Implicit Bias on Treatment Recommendations for 4 Common Pediatric Conditions: Pain, Urinary Tract Infection, Attention Deficit Hyperactivity Disorder, and Asthma. *American Journal of Public Health 102*(5), 988–995.

Sacarny, A., M. L. Barnett, J. Le, F. Tetkoski, D. Yokum, and S. Agrawal (2018, October). Effect of Peer Comparison Letters for High-Volume Primary Care Prescribers of Quetiapine in Older and Disabled Adults: A Randomized Clinical Trial. *JAMA Psychiatry 75*(10), 1003–1011.

Sacarny, A., D. Yokum, A. Finkelstein, and S. Agrawal (2016, March). Medicare Letters To Curb Overprescribing Of Controlled Substances Had No Detectable Effect On Providers. *Health Affairs 35*(3), 471–479.

Sacks, D. W., A. Hollingsworth, T. Nguyen, and K. Simon (2021, March). Can policy affect initiation of addictive substance use? Evidence from opioid prescribing. *Journal of Health Economics 76*, 102397.

Schnell, M. and J. Currie (2018, August). Addressing the Opioid Epidemic: Is There a Role for Physician Education? *American Journal of Health Economics 4*(3), 383–410.

Shapiro, B. T. (2018, December). Informational Shocks, Off-Label Prescribing, and the Effects of Physician Detailing. *Management Science 64*(12), 5925–5945.

Shurtz, I., Y. Goldstein, and G. Chodick (2024, January). Realization of Low-Probability Clinical Risks and Physician Behavior: Evidence from Primary Care Physicians. *American Journal of Health Economics 10*(1), 132–157.

Silver, D. (2021, May). Haste or Waste? Peer Pressure and Productivity in the Emergency Department. *The Review of Economic Studies 88*(3), 1385–1417.

Simeonova, E., N. Skipper, and P. R. Thingholm (2024, May). Physician Health Management Skills and Patient Outcomes. *Journal of Human Resources 59*(3), 777–809.

Simon, H. A. (1957). *Models of Man: Social and Rational; Mathematical Essays on Rational Human Behavior in Society Setting*. Wiley.

Singh, H., C. Spitzmueller, N. J. Petersen, M. K. Sawhney, and D. F. Sittig (2013, April). Information Overload and Missed Test Results in Electronic Health Record–Based Settings. *JAMA Internal Medicine 173*(8), 702–704.

Singh, M. and A. Venkataramani (2022, August). Rationing by Race. Working Paper 30380, National Bureau of Economic Research.

Sobczak, A., L. Taylor, S. Solomon, J. Ho, S. Kemper, B. Phillips, K. Jacobson, C. Castellano, A. Ring, B. Castellano, R. J. Jacobs, A. Sobczak, L. Taylor, S. Solomon, J. Ho, S. R. Kemper, B. Phillips, K. Jacobson, C. Castellano, A. Ring, B. Castellano, and R. J. Jacobs (2023, May). The Effect of Doulas on Maternal and Birth Outcomes: A Scoping Review. *Cureus 15*(5).

Sommers, B. D., C. L. McMurtry, R. J. Blendon, J. M. Benson, and J. M. Sayde (2017). Beyond Health Insurance: Remaining Disparities in US Health Care in the Post-ACA Era. *The Milbank Quarterly 95*(1), 43–69.

Stevenson, M. T. and J. L. Doleac (2022, September). Algorithmic Risk Assessment in the Hands of Humans. SSRN Scholarly Paper 3489440, Rochester, NY.

Tai-Seale, M. and T. McGuire (2012). Time is up: Increasing shadow price of time in primary-care office visits. *Health Economics 21*(4), 457–476.

van Parys, J. (2016, August). Variation in Physician Practice Styles within and across Emergency Departments. *PLOS ONE 11*(8).

Vela, M. B., A. I. Erondu, N. A. Smith, M. E. Peek, J. N. Woodruff, and M. H. Chin (2022, April). Eliminating Explicit and Implicit Biases in Health Care: Evidence and Research Needs. *Annual Review of Public Health 43*, 477–501.

Wallis, C. J. D., A. Jerath, N. Coburn, Z. Klaassen, A. N. Luckenbaugh, D. E. Magee, A. E. Hird, K. Armstrong, B. Ravi, N. F. Esnaola, J. C. A. Guzman, B. Bass, A. S. Detsky, and R. Satkunasivam (2022, February). Association of Surgeon-Patient Sex Concordance With Postoperative Outcomes. *JAMA Surgery 157*(2), 146–156.

Wilding, A., L. Munford, B. Guthrie, E. Kontopantelis, and M. Sutton (2022, September). Family doctor responses to changes in target stringency under financial incentives. *Journal of Health Economics 85*.

Williams, D. R., J. A. Lawrence, and B. A. Davis (2019, April). Racism and Health: Evidence and Needed Research. *Annual Review of Public Health 40*(Volume 40, 2019), 105–125.

Wilson, K. and L. Khansa (2018, November). Migrating to electronic health record systems: A comparative study between the United States and the United Kingdom. *Health Policy 122*(11), 1232–1239.

Wong, A., E. Otles, J. P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penoza, M. Ghous, and K. Singh (2021, August). External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine 181*(8), 1065–1070.

Wu, B. and G. David (2022, May). Information, relative skill, and technology abandonment. *Journal of Health Economics 83*, 102596.

Zeltzer, D., L. Einav, J. Rashba, and R. D. Balicer (2024, April). The Impact of Increased Access to Telemedicine. *Journal of the European Economic Association 22*(2), 712–750.

Zeltzer, D., L. Einav, J. Rashba, Y. Waisman, M. Haimi, and R. D. Balicer (2023, July). Adoption and utilization of device-assisted telemedicine. *Journal of Health Economics 90*, 102780.

Table 1: Variation in Physician Practice Style

Paper	Research Question	Data	Empirical Methods	Results	Heterogeneous Effects?
Abaluck et al. (AER 2016)	Variation in physician propensity to test for pulmonary embolism (PE) and effect of test misallocation on health outcomes.	20% sample Part B Medicare Claims 2000-2009; Part A claims with PE diagnosis; patient chart and billing data from two academic medical centers.	See text.	The average doctor tests if she believes the likelihood of a positive test is higher than 5.6 percent (SD = 5.4). Doctors react strongly to clinical symptoms but not to known PE risk factors from the patient's medical history.	
Ahammer and Schober (Health Economics, 2020)	How much of the variation in Austrian health expenditures is explained by GP practice style?	Upper Austrian Health Insurance Fund data (2005–2012); Medical Chamber data on doctor demographics; inpatient records.	AKM decomposition with patient and GP FEs, exploiting patients sho change GPs over time. Card et al. (2013) decomposition of variance.	Accounting for patient demand, patients of high-usage GPs have 20 to 148.5% higher expenditures than patients seeing an average GP.	Older doctors, female doctors, and doctors practicing in areas in higher GP density have higher expenditures.
Badinski et al. (NBER Working Paper, 2023)	How does geographic variation in physician practice intensity affect healthcare utilization?	20% random sample of Medicare fee-for-service claims 1998–2013.	Movers design exploiting patients and physician moves between HHRs and differences in utilization within HHRs estimated using patient and physician FE models.	A 1 SD increase in an HHR's average physician practice intensity increases utilization per visit 13%. 3/5 of the variation in an HHR's average physician practice intensity comes from variation within specialties and the rest from differences in physician specialty mix across HHRs.	Variation in PCP intensity across HHRs explains 19% of variation primary care utilization. Variation in cardiologist intensity explains only 3% of variation in cardiology utilization.
Berndt et al. (JHE 2015)	How concentrated are antipsychotic prescribing practices? (Do doctors have favorite drugs?)	10% sample from IMS retail prescriptions data, with refreshment each year; linked to the AMA Masterfile.	Descriptive.	Two thirds of a physician's prescriptions are for the same drug. The Herfindahl in prescribing concentration is decreasing in the log of total yearly antipsychotic prescriptions suggesting learning by doing.	The relationship between the volume of prescribing and the Herfindahl is larger for primary care physicians than for psychiatrists.

Chan, Gentzkow, and Yu (QJE, 2022)	Does radiologists' diagnostic skill affect diagnosis and outcomes for suspected pneumonia patients?	Veteran's Health Administration Emergency Department data Oct. 1999 to Sept. 2015.	See text.	Variation in skill explains 39% of the variation in diagnostic decisions and 78% of the variation in outcomes for suspected pneumonia patients. Diagnostic thresholds increase with skill.	
Currie, MacLeod and Van Parys (JHE, 2016)	Characterize practice style and describe how variation in practice style affects outcomes of heart attack patients?	Florida hospital discharge data for AMI patients admitted through the ED, 1992-2014; Data on providers from Florida medical license database.	Define appropriateness for invasive procedure using teaching programs. Regress use of invasive procedures on appropriateness and examine intercept (aggressiveness) and slope (responsiveness).	Within hospitals and years, patients with more aggressive providers have higher costs and better outcomes. Providers who follow "best practices" do too few procedures on healthy elderly suggesting over-reliance on age as a criterion.	Young, male providers from top schools are more aggressive.
Currie and MacLeod (JOLE, 2017)	How do variations in physician diagnostic and surgical skill affect outcomes of pregnancy?	~1 million NJ electronic birth records for 1997-2006.	See text.	Better diagnosis would reduce C-sections for low-risk mothers and increase C-sections for high-risk births, which would prevent infant death. Better surgical skills increase C-section rates and improve outcomes across the board.	Reducing C-section rates across the board would harm infants in high-risk pregnancies.
Cutler et al. (AEJ:EP 2019)	How does the percentage of "cowboys" and "comforters" in an area relate affect end-of-life spending.	Random sample of 598 cardiologists, 967 PCPs and 2,882 Medicare patients; Medicare expenditures from Dartmouth Atlas; Measures from the "Hospital Care" database.	Categorization of physicians based on survey results. Cowboys are physicians who recommend intensive care beyond current guidelines. Comforters recommend palliative care for the severely ill. Categories not mutually exclusive.	A 1 SD increase in the share of cowboys leads to 10.66-13.12% higher spending in last 2 years and a 2.15-3.56% higher 1-year spending for AMI patients. A 1 SD increase in the share of comforters leads to a 2.68-5.51% fall in spending in last 2 years, and a 0.82-1.2% fall in 1-year spending for AMI patients. Shares not significantly associated with survival.	
Fadlon and Van Parys (JHE 2020)	How does PCP practice style affect patient health care utilization?	20% sample of Medicare enrollees with at >=one month of traditional	Event study/d-in-d exploiting PCP changes when a patient's PCP relocates or retires.	Switching to a PCP whose patients spend \$10 more on primary care (PC) increases per capita spending 4.07%. Switching to a PCP whose patients have 1 SD more	Distinguish PCP switches within and between practices. Results similar

		Medicare enrollment in the year.		PC visits increases visits 38.20%. Similar effects for #diagnoses, flu vaccines, and diabetes care.	indicating variation is associated with individual PCPs.
Marquardt (WP, 2021)	How does physician practice style affect diagnosis of ADHD? What doctor characteristics predict practice style.	Electronic medical records from 129 doctors (12,311 pediatric patients) in a large healthcare system, Jan. 2014-Sept. 2017. Physician characteristics from the web.	Use natural language processing to measure child's suitability for ADHD diagnosis. Regress diagnosis on suitability. Examine intercept (intensiveness) and slope (compliance with guidelines). Regress doctor-specific estimates on doctor characteristics.	A physician with the median intensity (intercept) and median compliance (slope) diagnoses patients with the median symptom level 3.46% of the time. Increasing physician intensity by 1 SD increases diagnosis probability to 22.45%. Increasing physician compliance 1 SD increases diagnosis probability to 20.0%.	Less experienced male physicians have lower intercepts. Less experienced female physicians have higher slopes. Physicians who see patients with higher average severity have lower intensity and higher compliance.

Notes: See glossary for abbreviations.

Appendix for “Doctor Decision Making and Patient Outcomes”

Janet Currie[*], W Bentley MacLeod[†] and Kate Musen[‡]

February 16, 2025

[*] Princeton University and NBER
[†] Yale University and NBER
[‡] Columbia University

1 Appendix for Theory in Section 2.

This appendix lays out the detailed proofs of the model discussed in the text. The model begins with a population of patients where patient $i \in \mathscr{N}_j$ seeks treatment from doctor $j \in J$. It is assumed that neither patient or physician is sure which is the best choice. The doctor chooses between a non-intensive or an intensive treatment, denoted by $t_{ij} \in \{NI, I\}$. It is assumed that there is a best choice for the patient given by their *unobserved* state $\alpha_i \in \{L, H\}$. If $\alpha_i = L$, then the patient is low risk, and hence a non-intensive treatment is appropriate, while $\alpha_i = H$ implies that the patient is high risk, and an intensive treatment is more appropriate. This modeling strategy is based on Savage (1972 (first published 1954)'s model of Bayesian choice in which the goal of the model is not to provide a complete representation of the patient's condition, but to highlight only those aspects of a patient's state that are relevant for the decision at hand.[1]

Let the fraction of patients in $\mathscr{N}_j$ for which the doctors believe are low risk, $\alpha_i = L$, be given by $p_{Lj} \in (0,1)$, while a fraction $p_{Hj} = 1 - p_{Lj}$ the doctors suppose are in the high risk category, $\alpha_i = H$. Doctor j cannot perfectly observe the patient's state, but after examining the patient observes a signal:

$$T_{ij} = \begin{cases} 1 + \epsilon_i/\gamma_j, & \alpha_i = H, \\ -1 + \epsilon_i/\gamma_j, & \alpha_i = L, \end{cases} \tag{1}$$

where $\epsilon_i \sim N(0,1)$ and γ_j is the diagnostic skill of the doctor. An increase in diagnostic skill implies a more precise assessment of a person's state. The doctor is never perfectly sure of the patient's condition since it is observed with error.

T_{ij} is increasing with α_i so it follows that the doctor's decision criterion for the treatment choice $t_{ij} \in \{NI, I\}$ takes the form:

$$t_{ij} = \begin{cases} I, & T_{ij} \geq \tau_j, \\ NI, & T_{ij} < \tau_j, \end{cases}$$

where the doctor's decision threshold is given by τ_j.

The quality of diagnosis can be measured by the likelihood that a patient is assigned to the correct treatment. There are two measures of performance corresponding to whether patients correctly or incorrectly receive the intensive treatment. Suppose a patient is in state $\alpha_i = H$ and hence should be assigned to intensive treatment. The probability that the patient correctly receives the intensive treatment, given the doctor's decision threshold, τ_j, and diagnostic skill γ_j, the *true positive rate or TPR* is given by:

$$\begin{aligned} TPR(\tau_j, \gamma_j) &\equiv \Pr[T_{ij} \geq \tau_j | \alpha_i = H], \\ &= \Pr[1 + \epsilon/\gamma_j \geq \tau_j], \\ &= F(\gamma_j(1 - \tau_j)), \end{aligned} \tag{2}$$

where $F(\cdot)$ is the Normal cumulative probability distribution.

The probability that a patient who needs non-intensive treatment ($\alpha_i = L$) receives intensive treatment

[1]See the discussion in Chapter 2 of MacLeod (2022).

is given by the *false positive rate or FPR*:

$$\begin{aligned} FPR(\tau_j, \gamma_j) &\equiv \Pr[T_{ij} \geq \tau_j | \alpha = L] \\ &= \Pr[-1 + \epsilon/\gamma_j \geq \tau_j] \\ &= F(\gamma_j(-1 - \tau_j)). \end{aligned} \tag{3}$$

The Doctor's Decision Threshold (τ_j^*)

This section derives the doctor's decision threshold, τ_j^*, given a doctor's preferences and diagnostic skill, γ_j, and the consequences for a patient getting the inappropriate treatment. It is assumed that the doctor's utility is given by the well-being of the patient plus payments that might distort this decision. In particular, the doctor would make the socially efficient solution if their preferences are given by the patient utility less the cost of treatment. Given patient type $\alpha_i \in \{H, L\}$, doctor $j's$ utility from administering treatment $t \in \{NI, I\}$ is given by:

$$U_{\alpha tj} = u_{\alpha tj} + \delta_{tj}, \tag{4}$$

where $u_{\alpha tj}$ is the expected medical benefit to a patient of type $\alpha_i \in \{L, H\}$, getting treatment $t \in \{NI, N\}$ from doctor j. For the same patient type, the outcome $u_{\alpha tj}$ can differ by doctor, a variation that we associate with a doctor's *procedural skill*. Additional factors that affect treatment, such as a payment that the doctor receives from administering the treatment, are captured by δ_{tj}. We normalize this term by setting $\delta_{Lj} = 0$ and letting $\delta_j = \delta_{IJ} \in \Re$ be the pecuniary return (that can be positive or negative) from doing the intensive procedure.

For a type $\alpha_i = L$ patient a non-intensive treatment is preferred hence $u_{LNIj} > u_{LIj}$, while for type $\alpha_i = H$ intensive treatment is preferred and hence $u_{HIj} > u_{HNIj}$. If this were not the case, then there would be no diagnostic decision to make - all patients would be assigned to either intensive or non-intensive treatment. Let $\Delta_{HIj} = \{U_{HIj} - U_{HNIj}\}$ and $\Delta_{LNIj} = \{U_{LNIj} - U_{LIj}\}$ be the increase in utility for patients who receive the appropriate treatment. Notice that:

$$\begin{aligned} \Delta_{HIj} &= \{u_{HIj} - u_{HNIj}\} + \delta_{Ij}, \\ \Delta_{LNIj} &= \{u_{LNIj} - u_{LIj}\} - \delta_{Ij}. \end{aligned}$$

Hence we have the following lemma:

Lemma 1. *Regardless of the signal T_{ij}, when $\delta_{Ij} > u_{LNIj} - u_{LIj} > 0$ then the doctor j always provides the intensive treatment, and when $\delta_{IJ} < -\{u_{HIj} - u_{HNIj}\} < 0$, then the doctor always provides the non-intensive treatment.*

Proof. The proof follows from the fact that regardless of the information received, when $\delta_{Ij} > u_{LNIj} - u_{LIj} > 0$, then $\Delta_{LNIj} < 0$ and hence the doctor would choose the intensive treatment for the low type. This condition also implies that $\Delta_{HIj} > 0$, hence regardless of type, the intensive procedure is preferred. A similar argument applies when $\delta_{IJ} < -\{u_{HIj} - u_{HNIj}\} < 0$. □

This result points out that if the pecuniary returns for choice (δ_{It}) is either very positive or very negative, then the physician will always make the same treatment choice regardless of the signal. Thus in order to observe variation in treatment choice as a function of the doctor's information T_{ij}, the absolute value of pecuniary incentives cannot be too large. In the evidence we review, insensitivity to variation in observables may be due to either lack of an effect, or excess pecuniary returns.

The doctor's *ex ante* belief regarding the appropriate treatment for a patient in this pool of potential patients is given by:

$$p_{Hj} = \Pr\left[\alpha_i = H|j\right]$$

while the belief that the probability that $\alpha_i = L$ is $p_{Lj} = 1 - p_{Hj}$.

It is worth emphasizing that p_{Hj} is the doctor's subjective belief that may not necessarily equal the true probability, p_H. In general p_{Hj} is correlated with p_H, but there can be significant variation due to a number of doctor specific factors, including poor judgment and doctor biases.

The expected utility of doctor j who chooses decision threshold τ_j for patient i is given by:

$$\begin{aligned}
u_{ij}\left(\tau_j, \gamma_j\right) &= \left(\left(u_{HIj} + \delta_j\right)\Pr\left[T_{ij} \geq \tau_j|\alpha = H\right] + u_{HNI1}\Pr\left[T_{ij} < \tau_j|\alpha = H\right]\right)\Pr\left[\alpha = H|j\right] \\
&+ \left(\left(u_{LIj} + \delta_j\right)\Pr\left[T_{ij} \geq \tau_j|\alpha = L\right] + u_{LNIj}\Pr\left[T_{ij} < \tau_j|\alpha = L\right]\right)\Pr\left[\alpha = L|j\right] \\
&= \left(u_{HNIj} + \Delta_{HIj}\Pr\left[T_{ij} \geq \tau_j|\alpha = H\right]\right)p_{Hj} \\
&+ \left(u_{LIj} - \Delta_{LNIj}\Pr\left[T_{ij} \geq \tau_j|\alpha = L\right]\right)p_{Lj}, \\
&= u_j^0 + \Delta_{HIj}TPR\left(\tau_j, \gamma_j\right) \times p_{Hj} - \Delta_{LNIj}FPR\left(\tau_j, \gamma_j\right) \times p_{Lj},
\end{aligned} \tag{5}$$

where:

$$\begin{aligned}
u_j^0 &= u_{HNIj}\Pr\left[\alpha_i = H|j\right] + u_{LIj}\Pr\left[\alpha_i = L|j\right], \\
&= u_{HNIj} \times p_{Hj} + u_{LIj} \times p_{Lj}.
\end{aligned}$$

The quantity u_j^0 is the *worst* possible medical payoff for doctor j with any of their patients. It is the outcome when all individuals with type $\alpha = H$ are given the non-intensive treatment, and all type $\alpha = L$ individuals are given the intensive treatment. The payoff to a doctor can now be written in terms of the expected gains, beliefs and expected patient outcomes.

The decision threshold for each physician is $\tau_j^* = \arg\max_{\tau\in\Re} u_{ij}\left(\tau, \gamma_j\right)$. The solution is given by the following proposition.

Proposition 1. *The doctor's decision threshold solves $\tau_j^* = \arg\max_{\tau\in\Re} u_{ij}\left(\tau, \gamma_j\right)$. Suppose the pecuniary return satisfies $\delta_j \in \left(-\Delta_{HIj}, \Delta_{LNI}\right)$ (the conditions for lemma 1 are not satisfied), then τ_j^* satisfies the likelihood ratio condition:*

$$L\left(\tau_j^*, \gamma_j\right) = \frac{\Delta_{LNIj}}{\Delta_{HIj}} \times \frac{p_{Lj}}{p_{Hj}}, \tag{6}$$

where the likelihood ratio is given by:

$$L\left(\tau_j^*, \gamma_j\right) = \frac{f\left(\gamma_j\left(1 - \tau_j^*\right)\right)}{f\left(\gamma_j\left(-1 - \tau_j^*\right)\right)},$$

and $f\left(\cdot\right)$ is the Normal density function.

Proof. The solution satisfies the first order condition:

$$\begin{aligned}
0 &= \partial u_{ij}(\tau, \gamma_j)/\partial \tau, \\
&= (u_{HIj} + \delta_j)\,\partial TPR(\tau, \gamma_j)/\partial \tau \times p_{Hj} - \Delta_{LNIj}\partial FPR(\tau, \gamma_j)/\partial \tau \times p_{Lj}, \\
&= \Delta_{HIj} f(\gamma_j(1-\tau))(-\gamma_j) \times p_{Hj} - (\Delta_{LNIj} - \delta_j) f\left(\gamma_j\left(-1-\tau_j^*\right)\right)(-\gamma_j) \times p_{0j}.
\end{aligned}$$

The conditions on δ_j ensure that the ratio on the right of (6) is strictly positive. The first order condition follows from the last line. The first order conditions imply a unique decision threshold, τ_j^* satisfying:

$$L\left(\tau_j^*, \gamma_j\right) = \frac{f\left(\gamma_j\left(1-\tau_j^*\right)\right)}{f\left(\gamma_j\left(-1-\tau_j^*\right)\right)} = \frac{\Delta_{LNIj}}{\Delta_{HIj}} \times \frac{p_{Lj}}{p_{Hj}},$$

or:

$$\frac{\partial TPR(\tau, \gamma_j)/\partial \tau}{\partial FPR(\tau, \gamma_j)/\partial \tau} = \frac{\Delta_{LNIj}}{\Delta_{HIj}} \times \frac{p_{Lj}}{p_{Hj}}$$

When $\Delta_{HIj} < 0$ then $\Delta_{LNIj} > 0$ and doctor always does the non-intensive procedure. The converse holds when $\Delta_{LNIj} < 0$. □

The first order condition characterizes the global optimum, which follows from the Neyman-Pearson lemma showing that likelihood ratios are the most powerful form of hypothesis test (Neyman and Pearson (1933)).[2] When $\delta_j \in (-\Delta_{HIj}, \Delta_{LNI})$ the doctor faces uncertainty regarding choice. When this condition is not satisfied we say that the doctor is certain regarding her choice (either NI or I regardless of the test result). The model yields a closed form solution for the doctor's diagnostic rule τ_j^*, given by the following proposition:

Proposition 2. *When the doctor is uncertain, the decision threshold is given by:*

$$\tau_j^* = b_j^*/\gamma_j^2, \tag{7}$$

where $b_j^* \equiv \left(\ln\left(\Delta_{LNIj}/\Delta_{HIj}\right) + \ln\left(p_{Lj}/p_{Hj}\right)\right)/2$.

Proof. Observe:

$$\begin{aligned}
\frac{f\left(\gamma_j\left(1-\tau_j^*\right)\right)}{f\left(\gamma_j\left(-1-\tau_j^*\right)\right)} &= \frac{\exp - \left\{\gamma_j\left(1-\tau_j^*\right)\right\}^2/2}{\exp - \left\{\gamma_j\left(-1-\tau_j^*\right)\right\}^2/2} \\
&= \exp\left(-\left\{\gamma_j\left(1-\tau_j^*\right)\right\}^2 + \left\{\gamma_j\left(-1-\tau_j^*\right)\right\}^2\right)/2
\end{aligned}$$

[2]Feng et al. (2023) highlight the link between rational choice and the Neyman-Pearson lemma.

Taking the logarithm of the first-order condition gives us:

$$\left(-\left\{\gamma_j\left(1-\tau_j^*\right)\right\}^2+\left\{\gamma_j\left(-1-\tau_j^*\right)\right\}^2\right)/2=2\times b_j,$$
$$\left(-\left\{\gamma_j^2\left(1-2\tau_j^*+\left(\tau_j^{*2}\right)^2\right)\right\}+\gamma_j^2\left(1+2\tau_j^*+\left(\tau_j^{*2}\right)^2\right)\right)=4b_j$$
$$4\gamma_{ij}^2\tau_{ij}=4b_j,$$

giving the desired result (7). □

Equation (7) shows that the doctor's decision threshold depends on diagnostic skill, γ_j, the relative desirability of non-intensive and intensive treatments for the two types of patients, $\Delta_{LNIj}/\Delta_{HIj}$, and the doctor's beliefs about the relative proportions of patient types, p_{Lj}/p_{Hj}, in the population. When the doctor believes that there is a higher probability that the patient needs non-intensive treatment, she adopts a higher threshold resulting in less use of the intensive treatment. Similarly, if the relative benefit from intensive treatment is higher, then this results in a lower threshold.

As diagnostic skill increases, both patient types are more likely to be allocated to the appropriate treatment. The doctor's decision rule entails patients getting the appropriate treatment with probability close to one as diagnostic skill increases. Conversely, as diagnostic skill falls, the b_j term dominates. When $b_j>0$, treatment is biased in favor of the non-intensive treatment and the probability that patients are treated with the non-intensive procedure rises as diagnostic skill falls. When $b_j<0$, treatment is biased in favor of intensive treatment and the probability of intensive treatment rises as diagnostic skill falls. In effect, as diagnostic skill falls, physicians choose the treatment that they believe is best for most of their patients. These observations are summarized in the following proposition:

Proposition 3. *For a doctor who is uncertain of the best course of action (b_{ij} is finite), then as diagnostic skill increases, each patient is more likely to receive treatment appropriate for their type. More precisely:*

$$\lim_{\gamma_j\to\infty}\tau_j^*=1/2,$$
$$\lim_{\gamma_j\to\infty}u_{ij}^*=\begin{cases}u_{HIj}, & \text{if } \alpha_i=H,\\ u_{LNIj}, & \text{if } \alpha_i=L.\end{cases}$$

As diagnostic skill falls, all patients get the same treatment depending upon the sign of the decision shifter, b_j:

$$\lim_{\gamma_j \to 0} \tau_j^* = \begin{cases} \infty, & \text{if } b_j > 0, \\ 1/2, & \text{if } b_j = 0 \\ -\infty, & \text{if } b_j < 0. \end{cases}$$

$$\lim_{\gamma_j \to 0} u_{ij}^* = \begin{cases} u_{HNIj}, & \text{if } \alpha_i = H, b_j > 0, \\ u_{LNIj}, & \text{if } \alpha_i = L, b_j > 0, \\ \left(u_{HNIj} + u_{HIj}\right)/2, & \text{if } \alpha_i = H, b_j = 0, \\ \left(u_{LNIj} + u_{LIj}\right)/2, & \text{if } \alpha_i = L, b_j = 0, \\ u_{HIj}, & \text{if } \alpha_i = H, b_j < 0, \\ u_{LIj}, & \text{if } \alpha_i = L, b_j < 0. \end{cases}$$

Proof. The proof of this proposition follows from equation (7). □

1.1 Identifying the Doctor Diagnostic threshold, Diagnostic Skill, and Procedural Skill From Data

Proposition 4. *Given points (TPR_j, FPR_j) on an ROC curve generated by Normal errors, there is a unique solution (τ_j, γ_j) to:*

$$TPR_j = F\left(\gamma_j\left(1 - \tau_j\right)\right),$$
$$FPR_j = F\left(-\gamma_j \tau_j\right).$$

Proof. Since $(TPR_j, FPR_j) \in (0,1)^2$, we have:

$$\gamma_j\left(1 - \tau_j\right) = F^{-1}\left(TPR_j\right), \tag{8}$$
$$-\gamma_j \tau_j = F^{-1}\left(FPR_j\right). \tag{9}$$

Plugging (9) into (8) we get:

$$\begin{aligned} \gamma_j\left(1 - \tau_j\right) &= \gamma_j - \gamma_j \tau_j, \\ &= \gamma_j + F^{-1}\left(FPR_j\right), \end{aligned}$$

and hence:

$$\gamma_j = F^{-1}\left(TPR_j\right) - F^{-1}\left(FPR_j\right).$$

It must be the case that $\gamma_j > 0$ since from the properties of ROC curves we have $TPR_j - FPR_j > 0$ and the fact that the cumulative distribution function $F()$ is strictly increasing. Using (9) we get:

$$\tau_j = -F^{-1}\left(FPR_j\right)/\gamma_j.$$

□

Abaluck et al. (2016)

The context for Abaluck et al. (2016) is ordering computerized tomography (CT) scans to test for a pulmonary embolism (PE). The use of scans is expensive, and while a pulmonary embolism is a serious condition. The goal of the paper is to ask whether or not there is excessive use of CT scans? In the context of our model, a CT scan is an intensive procedure, hence $t_{ij} = I$ if a doctor j orders a scan for patient i. The unobserved state is whether a person has a PE ($\alpha_i = H$), or does not ($\alpha_i = L$). The goal is to have a true positive rate of 1, which ensures that all individuals with a PE are tested and treated. However, the test is expensive and it is not always possible for the doctor to correctly assess the patient's condition. In general one expects to have a TPR < 1 and a FPR > 0.

The goal of the paper is to assess the extent to which the decision threshold varies between doctors, and the extent to which doctors process information correctly. The challenge is that, unlike Chan, Gentzkow and Yu (2022), patients are not randomly allocated to doctors, and hence the average severity of the cases can vary by doctor. The authors address this by specifying and estimating a structural model of physician decision making. It is assumed that the signal on the condition of patient i is the expected probability that has a PE:

$$\begin{aligned} T_{ij} &= \Pr[\alpha = H|i, j] && (10) \\ &= \vec{x}_i\beta + a_j + \eta_{ij}, && (11) \\ &\equiv \rho_j(\vec{x}_i) + \eta_{ij} && (12) \end{aligned}$$

where η_{ij} is information observed by the doctor, but not the econometrician, and $\rho_j(\vec{x}_i) = \Pr[\alpha_i = H|\vec{x}_i, j]$ is the probability that the individual has PE conditional upon the observables $\vec{x}_i$ and the population of patients treated by docter j.

In this case, the decision threshold, τ_j^*, defines the cutoff probability for ordering a CT-scan. When the probability of a PE is greater than τ_{j*} then the doctor orders a CT-scan.

A key feature of this specification is the inclusion of the fixed effect a_j that captures the fact that doctors may face different distributions of patients. If patients were randomly allocated, then $a_j = a$ for some constant a for all doctors. We shall show that the challenge will be to separately estimate both a_j and the doctor's decision threshold τ_{j*}.

The authors suppose that the distribution of η_{ij} is a known *i.i.d.* distribution that is independent of patient observables $\vec{x}_i$, and with distribution $\eta_{ij} \sim H(\cdot)$, where $H(\eta) \equiv \Pr[\eta_{ij} \leq \eta]$ is the cumulative probability distribution. It is assumed $E\{\eta_{ij}\} = 0$. The online appendix of Abaluck et al. (2016) provides a parametric specification for $H(\cdot)$ and it is shown that it can be estimated from the data. For the current discussion, it is assumed to be known.

Given the single index T_{ij}, Abaluck et al. (2016) and doctor practice style characterized by a threshold τ_j^*, a test is ordered whenever it is suspected that the probability of a PE is greater than τ_j^*:

$$t_{ij} = \begin{cases} I, & T_{ij} \geq \tau_j^*, \\ NI, & T_{ij} \leq \tau_j^*, \end{cases}$$

Thus, doctor j orders a test if and only if:

$$\begin{aligned} T_{ij} - \tau_j^* &\geq 0, \\ \vec{x}_i \beta + a_j - \tau_j^* + \eta_{ij} &\geq 0, \\ \vec{x}_i \beta + \hat{a}_j + \eta_{ij} &\geq 0. \end{aligned}$$

Thus, the probability a test is ordered is given by:

$$\begin{aligned} \Pr\left[t_{ij} = I | \vec{x}_i, j\right] &= \Pr\left[T_{ij} \geq \tau_j^* | \vec{x}_i, j\right] \\ &= \Pr\left[\rho_j\left(\vec{x}_i\right) + \eta_{ij} \geq \tau_j^* | \vec{x}_i, j\right] \\ &= \Pr\left[\eta_{ij} \geq \tau_j^* - \rho_j\left(\vec{x}_i\right) | \vec{x}_i, j\right] \\ &= 1 - H\left(\vec{x}_i \beta + \hat{a}_j\right). \end{aligned} \tag{13}$$

When estimating (13) it is not possible to separately identify τ_j^* and a_j. Rather, one can use (13) to estimate the intercept term $\hat{a}_j \equiv a_j - \tau_j^*$ and the coefficients β and whether or not a person has PE.

To estimate τ_j^* one needs information on the probability of a PE. From the above estimate, we can define:

$$\begin{aligned} \mathrm{s}_j\left(\vec{x}_i\right) &= \rho_j\left(\vec{x}_i\right) - \tau_j^*. \\ &= \left(\vec{x}_i \beta + a_j - \tau_j^*\right) \\ &= \left(\vec{x}_i \beta + \hat{a}_j\right) \end{aligned}$$

This function can be estimated from the data using (13), and the fact that the distribution of η_{ij} is known. The expected PE for tested individuals uses (10) to get:

$$\begin{aligned} \Pr\left[\alpha_i = H | \vec{x}_i, t_{ij} = I\right] &= \vec{x}_i \beta + a_j + E\left[\eta_{ij} | \vec{x}_i, t_{ij} = I\right] \\ &= \vec{x}_i \beta + a_j + E\left[\eta_{ij} | \eta_{ij} \geq \tau_j - \rho_j\left(\vec{x}_i\right)\right] \\ &= \tau_j^* + s_j\left(\vec{x}_i\right) + \int_{-s_j(\vec{x}_i)}^{\infty} \eta h\left(\eta\right) d\eta / \left(1 - H\left(\tau_j - \rho_j\left(\vec{x}_i\right)\right)\right), \\ &\equiv \tau_j^* + \lambda\left(s_j\left(\vec{x}_i\right)\right). \end{aligned} \tag{14}$$

where $h\left(\eta\right) = H'\left(\eta\right)$.[3] The key observation made by Abaluck et al. (2016) is that by construction it must be the case that $\Pr\left[\alpha_i = H | \vec{x}_i, t_{ij} = I\right] \geq \tau_j^*$, the cutoff probability. Under the hypothesis that some patients are not tested because the probability of PE is less than τ_j, implies that there exist marginal patients for which $\Pr\left[\alpha_i = H | \vec{x}_i, t_{ij} = I\right] = \tau_j^*$. The marginal patients are defined by:

$$M_j = \left\{i | \lambda\left(s\left(\vec{x}_i\right)\right) \approx 0, t_{ij} = I\right\}.$$

[3]Abaluck et al. (2016) allows for an error term with mass point. One simply adjusts the definition of the integral to allows for such mass points, which formally is the requirement that $H\left(s\right)$ is right continuous, with jumps at the mass points.

When the number of marginal patients is sufficiently large, then we can obtain an estimate of τ_j from:

$$\tau_j^* \simeq \frac{\sum_{i\in M_j} I_{\alpha_i=H}}{|M_i|}, \tag{15}$$

where $|M_j|$ is the number of patients in the marginal set, and $I_{\alpha_i=H} = 1$ when is $\alpha_i = H$ and zero otherwise. The implicit assumption is that the result from the CT scan is definitive and hence the true α_i is known for tested individuals. When this set M_j is large enough the authors are able to get a precise estimate of doctor's decision threshold or practice style. They show that the decision threshold does vary between doctors.

Computing the TPR and FPR

Finally, within this framework one can map the decision threshold, τ_j, into the ROC model as used by Chan, Gentzkow and Yu (2019). Here we rely upon the structural estimates for β,a_j and the distribution $H(\cdot)$. The unconditional probability a person with condition $\vec{x}_i$ has a PE is given by:

$$\rho_j(\vec{x}_i) \equiv \vec{x}_i\beta + a_j \in [0,1].$$

Thus, given that for each doctor a_j is known, then we can write the probability of persons tested having a PE from (14) as a function of potential decision threshold, τ_j, as:

$$\begin{aligned}
\Pr[\alpha_i = H|t_{ij} = I, \vec{x}_i, j, \tau_j] &= \rho_j(\vec{x}_i) + E[\eta_{ij}|\rho_j(\vec{x}_i) + \eta_{ij} \geq \tau_j] \\
&= \rho_j(\vec{x}_i) + \int_{\tau_j-\rho_j(\vec{x}_i)}^{\infty} \eta_{ij} h(\eta)\, ds/\left(1 - H(\tau_j - \rho_j(\vec{x}_i))\right), \\
&= \rho_j(\vec{x}_i) + \hat{\eta}(\tau_j - \rho_j(\vec{x}_i))/\left(1 - H(\tau_j - \rho_j(\vec{x}_i))\right),
\end{aligned}$$

where

$$\hat{\eta}(s) \equiv \int_s^{\infty} \eta h(\eta)\, ds,$$

is the mean value of the unobserved term, η_{ij}, greater than s. Since the mean of $\eta_{ij} = 0$ then is must be the case that $\hat{\eta}(s) \geq 0$. The support of η_{ij} must be finite in order for T_{ij} defined in (10) to be a probability, and hence $\hat{\eta}(s) = 0$ for $s > \bar{s}$ for some $\bar{s}$. From these we can compute the TPR and FPR for this model using Bayes rule:

$$\begin{aligned}
TPR(\vec{x}_i, a_j, \tau_j) &\equiv \Pr[t_{ij} = I|\alpha_i = H, \vec{x}_i, a_j, \tau_j] \\
&= \Pr[\alpha_i = H|t_{ij} = I, \vec{x}_i, a_j, \tau_j] \times \frac{\Pr[t_{ij} = I|\vec{x}_j, a_j, \tau_j]}{\Pr[\alpha_i = 1|\vec{x}_j, a_j]} \\
&= \left(\rho_j(\vec{x}_i) + \frac{\hat{\eta}(\tau_j - \rho_j(\vec{x}_i))}{(1 - H(\tau_j - \rho_j(\vec{x}_i)))}\right) \frac{(1 - H(\tau_j - \rho_j(\vec{x}_i)))}{\rho_j(\vec{x}_i)} \\
&= \left(1 - H(\tau_j - \rho_j(\vec{x}_i)) + \frac{\hat{\eta}(\tau_j - \rho_j(\vec{x}_i))}{\rho_j(\vec{x}_i)}\right).
\end{aligned}$$

To compute the corresponding FPR, using Bayes rule we get:

$$\Pr\left[t_{ij}=I|\vec{x}_j,a_j,\tau_j\right]=$$
$$FPR\left(\vec{x}_i,a_j,\tau_j\right)\times\Pr\left[\alpha_i=L|\vec{x}_j,a_j\right]+TPR\left(\vec{x}_i,a_j,\tau_j\right)\times\Pr\left[\alpha_i=H|\vec{x}_j,a_j\right]$$

From this we get:

$$\begin{aligned}FPR\left(\vec{x}_i,a_j,\tau_j\right)&=\frac{1-H\left(\tau_j-\rho_j\left(\vec{x}_i\right)\right)-TPR\left(\vec{x}_i,a_j,\tau_j\right)\times\rho_j\left(\vec{x}_i\right)}{1-\rho_j\left(\vec{x}_i\right)}\\&=1-H\left(\tau_j-\rho_j\left(\vec{x}_i\right)\right)-\frac{\hat{\eta}\left(\tau_j-\rho_j\left(\vec{x}_i\right)\right)}{1-\rho_j\left(\vec{x}_i\right)}\end{aligned}$$

We can see the shape of the ROC curve by looking at:

$$\begin{aligned}\Delta\left(\vec{x}_i,a_j,\tau_j\right)&=TPR\left(\vec{x}_i,a_j,\tau_j\right)-FPR\left(\vec{x}_i,a_j,\tau_j\right),\\&=\hat{\eta}\left(\tau_j-\rho_j\left(\vec{x}_i\right)\right)\left(\frac{1}{\rho_j\left(\vec{x}_i\right)}+\frac{1}{1-\rho_j\left(\vec{x}_i\right)}\right),\\&=\frac{\hat{\eta}\left(\tau_j-\rho_j\left(\vec{x}_i\right)\right)}{\rho_j\left(\vec{x}_i\right)\left(1-\rho_j\left(\vec{x}_i\right)\right)}.\end{aligned}$$

Hence the ROC curve can be parameterized via τ_j and given by:

$$TPR\left(\vec{x}_i,a_j,\tau_j\right)=\frac{\hat{\eta}\left(\tau_j-\rho_j\left(\vec{x}_i\right)\right)}{\rho_j\left(\vec{x}_i\right)\left(1-\rho_j\left(\vec{x}_i\right)\right)}+FPR\left(\vec{x}_i,a_j,\tau_j\right). \tag{16}$$

Currie and MacLeod (2017)

This paper uses the model outlined above, where T_{ij} is a signal of patient appropriateness for an intensive procedure (a C-section). From observational data, one observes the doctor's treatment choice ($t_{ij}\in\{NI,I\}$), and some measure of patient outcomes following treatment, as well as some information on patient type that may be available in medical records. Let $\vec{x}_i$ be patient characteristics that are observable in the data. Currie and MacLeod (2017) use the vector of observed patient characteristics, $\vec{x}_i$, to estimate the probability that $\alpha_i=H$, denoted by $\rho\left(\vec{x}_i\right)=\Pr\left[\alpha_i=H|\vec{x}_i\right]$. This is estimated using the full population of patients in New Jersey, and hence it provides a measure of appropriateness that is independent of physician characteristics and practice style.

It is assumed that each physician chooses τ_j^*, as derived in the model section. This in turn determines the TRP_j and FPR_j for the doctor. Here one is implicitly assuming that the signal T_{ij} has the information contained in $\vec{x}_i$. With this definition we have:

Proposition 5. *The doctor's estimated likelihood of performing an intensive procedure is:*

$$\Pr\left[t_{ij}=I|j,\vec{x}_i\right]=\left(TPR_j-FPR_j\right)\rho\left(\vec{x}_i\right)+FPR_j, \tag{17}$$

where $\rho\left(\vec{x}_i\right)=\Pr\left[\alpha_i=H|\vec{x}_i\right]$ is the estimated probability that the patient needs an intensive intervention, while TPR_j and FPR_j are computed at the doctor's decision rule (proposition 2). The slope term, $\theta_j=$

$(TPR_j - FPR_j)$ is increasing with a doctor's diagnostic skill:

$$\frac{d\theta_j}{d\gamma_j} > 0.$$

Finally, $\frac{d\theta_j}{db_j} > 0$ for $b_j < 0$ and $\frac{d\theta_j}{db_j} < 0$ for $b_j > 0$, namely the treatment decision is most sensitive to the prior condition of the patient $(\rho(\vec{x}_i))$ when $b_j^* = 0$.

Proof. The probability of a C-section is:

$$\begin{aligned}
\Pr[t_{ij} = I | j, \vec{x}_i] &= \Pr[t_{ij} = I | \alpha_i = H, \vec{x}_i, a_j, \tau_j] \times \Pr[\alpha_i = H | j, \vec{x}_i] \\
&\quad + \Pr[t_{ij} = I | \alpha_i = L, \vec{x}_i, a_j, \tau_j] \times \Pr[\alpha_i = L | j, \vec{x}_i] \\
&= TPR_j \times \rho_j(\vec{x}_i) + FPR_j \times (1 - \rho_j(\vec{x}_i)), \\
&= (TPR_j - FPR_j)\rho_j(\vec{x}_i) + FPR_j.
\end{aligned}$$

Then we have using the decision rule from proposition (1):

$$\begin{aligned}
\frac{d\theta_j}{d\gamma_j} &= \frac{dF\left(\gamma_j\left(1 - \tau_j^*\right)\right)}{d\gamma_j} - \frac{dF\left(\gamma_j\left(-1 - \tau_j^*\right)\right)}{d\gamma_j} \\
&= \frac{dF\left(\gamma_j - b_j^*/\gamma_j\right)}{d\gamma_j} - \frac{dF\left(-\gamma_j - b_j^*/\gamma_j\right)}{d\gamma_j} \\
&= \frac{b_j}{\gamma_j^2}\left(f\left(\gamma_j - b_j^*/\gamma_j\right) - f\left(-\gamma_j - b_j^*/\gamma_j\right)\right) \\
&= \frac{b_j}{\gamma_j^2}\exp\left(\gamma_j^2 + \frac{b_j^*}{\gamma_j^2}\right)\left(\exp\left(b_j^*\right) - \exp\left(-b_j^*\right)\right).
\end{aligned}$$

When $b_j > 0$ then $(\exp(b_j) - \exp(-b_j)) > 0$ and when $b_j < 0$, then $(\exp(b_j) - \exp(-b_j)) < 0$, Hence the right hand side is strictly positive when $b_j \neq 0$ and zero when $b_j = 0$, Thus the slope increases with skill.

In the case of b_j we have:

$$\begin{aligned}
\frac{d\theta_j}{db_j} &= \frac{dF\left(\gamma_j\left(1 - \tau_j^*\right)\right)}{db_j} - \frac{dF\left(\gamma_j\left(-1 - \tau_j^*\right)\right)}{db_j} \\
&= \frac{dF(\gamma_j - b_j/\gamma_j)}{db_j} - \frac{dF(-\gamma_j - b_j/\gamma_j)}{db_j} \\
&= -\frac{1}{\gamma_j}\left(f(\gamma_j - b_j/\gamma_j) - f(-\gamma_j - b_j/\gamma_j)\right) \\
&= -\frac{1}{\gamma_j}\exp\left(\gamma_j^2 + \frac{b_j}{\gamma_j^2}\right)\left(\exp(b_j) - \exp(-b_j)\right).
\end{aligned}$$

Hence, θ_j increases with b_j if and only if $b_j < 0$. Thus θ_j is largest when $b_j = 0$, and given by:

$$\theta_j \leq F(\gamma_j) - F(-\gamma_j)$$

□

Notice that from equation (17), as long as there is sufficient variation in the likelihood of needing intensive treatment, $\rho(\vec{x}_i)$, one can separately identify TPR_j and FPR_j in equation (17) Hence we can identify both τ_j and γ_j.

The slope term is also affected by the physician's beliefs about when invasive procedures are likely to be warranted via τ_j, and by any additional physician-specific factors that are included in δ_j. Currie and MacLeod (2017) distinguish between τ_j and γ_j by noting that in a doctor-specific regression, the constant term in Equation (17) is affected only by τ_j so given two estimated parameters and two unknowns, it is possible to identify both.

Finally, notice that patients with high *ex ante* likelihood of having a C-section ($\rho(\vec{x}_i) \approx 1$) then variation in patient outcomes is independent of both diagnostic skill and the decision threshold. Hence, we can associate variation in outcomes with procedural skill. A similar implication follows for patients with a low likelihood of a C-section ($\rho(\vec{x}_i) \approx 0$).

References

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh. 2016. "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care." *American Economic Review*, 106(12): 3730–3764.

Chan, David C, Jr, Matthew Gentzkow, and Chuan Yu. 2019. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." National Bureau of Economic Research Working Paper 26467.

Chan, David C, Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." *The Quarterly Journal of Economics*, 137(2): 729–783.

Currie, Janet M., and W. Bentley MacLeod. 2017. "Diagnosis and Unnecessary Procedure Use: Evidence from C-section." *Journal of Labor Economics*, 35(1): 1–42.

Feng, Kai, Han Hong, Ke Tang, and Jingyuan Wang. 2023. "Statistical Tests for Replacing Human Decision Makers with Algorithms."

MacLeod, W Bentley. 2022. *Advanced Microenomics for Contract, Instituional and Organizational Economics.* Boston, MA:MIT Press.

Neyman, Jerzy, and Egon Sharpe Pearson. 1933. "IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706): 289–337.

Savage, Leonard J. 1972 (first published 1954). *The Foundations of Statistics.* New York, N.Y.:Dover Publications.

Appendix Describing Research Papers Organized by Topic

Appendix Table 1: Health Disparities

Paper	Research Question	Data	Empirical Methods	Results	Heterogeneous Effects?
Alsan, Garrick, and Graziani (AER 2019)	How does physician race affect Black men's take up of preventative care services?	Experimental data with 1,374 recruited Black male participants, with 637 completing the study	Field experiment with random assignment to either a Black or non-Black physician in a special clinic offering preventive care. Doctor race was signaled to patients by a headshot.	Viewing the headshot did not significantly affect intended take-up of services. But patients who saw a Black patient increased demand for services ex-post by 38.79% for diabetes screening, 52.77% for cholesterol screening and 26.54% for flu shots.	No differences by income, education, or age. Effects greater for patients without a recent medical screening, with more ER visits, and with higher levels of measured medical mistrust.
Angerer, Waibel, and Stummer (AJHE 2019)	What is the effect of socioeconomic status, signaled by education level, on the probability of receiving a medical appointment and on response times?	Experimental data for April 26-June 2, 2017, with email requests for appointments sent to 1,249 Austrian specialists.	Correspondence study via email with varying email signatures to signal no degree, a doctoral degree, or a medical degree	Patients with degrees are more likely to receive an appointment, and have lower response times and lower waiting times. Whether patients are offered an appointment depends on the assistant, while response and waiting times depend on the doctor.	The effects are driven by practices that do not contract with social insurance.
Button et al. (NBER WP 2020)	How does being nonbinary or transgender interact with patient race to affect the probability of getting an appointment with a mental health care provider (MHP)?	Experimental correspondence data from 1,000 emails sent to MHPs between Jan. 28, 2020-May 15, 2020, with number of emails per zip code proportional to population.	Emails sent through an MHP appointment request website with randomly assigned content disclosing trans or nonbinary status. Names signal gender and race. Randomize whether help is sought for depression, anxiety, or "stress."	Transgender or non-binary African Americans and Hispanics are 18.7% less likely to get a positive response than cisgender whites. No evidence of differential responses by TNB status for whites.	N/A
Brekke et al. (HE 2018)	What is the relationship between SES of Type II diabetes patients and GP treatment decisions?	Norwegian administrative health data 2008- 2012; patient and GP characteristics from Statistics Norway.	GP FE models of service provision conditional on patient characteristics. Additional results using GP quits, retirements, and moves.	High ed. patients get fewer, longer visits, Less ed. patients get more medical tests and services over the course of a year. E.g. high ed. 14.79% more likely to get a visit over 20 minutes. Less ed. 3.94% more likely to get 2+ HbA1C tests.	Results are similar when disaggregated by patient age and GP sex, age, specialty, number of patients, and fixed payment vs. fee-for-service.
Cabral and Dillender (AER	How does gender concordance between	Open records request for Texas worker's	Assignment to doctors is random conditional on	Female claimants seen by a female doctor are 5.2% more likely to receive	Differences are not statistically significant but

2024)	claimants and doctors performing independent medical evaluations for workers compensation affect disability determinations?	compensation claims 2013-17, and independent medical evaluations 2005-2017; NPI registry; novel survey of 1,519 adults 30-64, 2021.	doctor's credential and the claimants' county. Estimate OLS with an interaction between female doctor and female claimant controlling for main effects, credential, and county.	benefits compared to when female claimants are seen by male doctors. Physician gender does not affect likelihood of receiving benefits for male claimants. Female claimants seen by a female doctor receive 8.6% higher benefits than female claimants seen by male doctors.	suggest larger effects for those with lower earnings, in less dangerous industries, but with worse injuries.
Chandra and Staiger (NBER WP 2010)	Are differences in the treatment of Black and female AMI patients due to physician preferences or statistical discrimination?	Clinical records for 200,000+ patients admitted for AMI in 1994 & 1995 from the Cooperative Cardiovascular Project (CCP).	Propensity score estimation; taste based discrimination implies that similar patients who receive fewer services will suffer worse outcomes.	Black and female patients receive less treatment but also receive slightly lower benefits from treatment suggesting that they are not being denied beneficial treatment due to discrimination.	N/A.
Eli, Logan, and Miloucheva (NBER WP 2019)	Use union army pension awards to examine the effect of income on mortality. Investigate differences in a board's disability evaluations by race of applicant.	Union Army and United States Colored Troops (USCT) sample from the Early Indicators Project; Rosters of Examining Surgeons from the National Archives.	Instrument pension income using leave-one-out mean of a board's pension determinations. Include board FEs. First stage shows the same boards were less generous to Black veterans.	Pension income significantly increased life expectancy. Bias against Black veterans in determining pension eligibility is substantial and accounts for much of the racial mortality gap in this population.	Bias against Black veterans is strongest for conditions where valuations may be more subjective, such as digestive diseases.
Frakes and Gruber (NBER WP 2022)	How does the availability of Black physicians on a military base affect Black Tricare patients' outcomes?	Military Health System Data Repository fiscal years 2003–2013	Mover-based ITT design exploiting differences in racial shares of physicians across bases.	1 SD increase in share of Black physicians reduces Black patients' mortality from diabetes, hypertension, high cholesterol, and cardiovascular disease by 15%. 55–69% of the effect attributed to medication adherence.	N/A.
Goyal et al. (JAMA Pediatrics 2015)	How does treatment of pain in the ED vary by race for child appendicitis patients?	National Hospital Ambulatory Medical Care Survey 2003-2010.	Multivariate logistic regression.	Black patients were less likely to receive any analgesia, adjusted OR=0.1 for moderate pain and 0.2 for severe pain. Black patients were less likely to receive opioids, adjusted OR= 0.2.	The authors test for interactions between race and sex but do not find any.

Greenwood, Carnahan, and Huang (PNAS 2018)	How does patient-attending gender concordance affect mortality from heart attacks among patients admitted to the ED? Do male doctors with more female colleagues or AMI patients have better female survival?	Census of patients admitted to hospitals in Florida 1991- 2010 from Florida's Agency for Healthcare Administration.	Assume patient assignment to physicians is conditionally random in the ED and either include physician FEs or hospital-quarter FEs. They also estimate additional specifications using matching.	In the full sample with hospital-quarter FEs, relative to male or female patients treated by female physicians, female patients treated by male doctors are 1.80% less likely to survive and male patients treated by male doctors are 0.90% less likely to survive. In the matched sample, only female patients treated by male doctors have lower survival rates.	Female survival increases when there are more female physicians in the ED, especially when they are treated by male physicians. Female patients treated by male physicians are more likely to survive as the number of female patients their doctor has treated in the prior quarter increases.
Greenwood et al. (PNAS 2020)	How does infant and maternal mortality vary as a function of patient-doctor racial concordance?	Census of patients admitted to hospitals in Florida 1992- 2015 from Florida's Agency for Healthcare Administration.	OLS with controls including physician FEs in some models.	Racial concordance between infant and physician corresponds to about a 40% reduction in gap in mortality between Black and white infants. No significant racial concordance effects are found for mothers.	Effects are more precisely estimated for infants with >=1 comorbidity and for infants in hospitals that see more Black patients. Effects are similar in % terms for pediatricians and non-pediatricians.
Hill, Jones, and Woodworth (JHE 2023)	What is the effect of physician-patient race concordance on within-hospital mortality among uninsured non-Hispanic, Black and white patients admitted through the ED?	Florida Hospital Discharge Data File from October 2011 to December 2014; Florida Physician Workforce Survey from 2008-2016.	IV measures "the lagged share of same-race physicians typically present at the indexed hospital on the weekday and shift" when patient admitted.	Physician-patient race concordance reduces mortality by 27%.	The largest effects are for subgroups of patients with high variance in number of procedures and in total charges.
Hoffman et al. (PNAS 2016)	How do false beliefs about biological racial differences among white doctors mediate racial differences in recommended for hypothetical patients?	Experimental and survey data from U.S. medical students and residents (N=222 after restricting to white, US-born, native English-speaking).	Surveys and experimental vignettes.	Participants one SD above the mean in terms of false beliefs rated the Black patient as having 0.45 less pain than the white patient on a scale of 1-10 and were less accurate in recommendations for the Black patients.	Some statistics are disaggregated by medical school year or resident status, but sample sizes are too small to draw inferences.
McDevitt and Roberts (RAND	How does the availability of female	American Medical Information's data on	Descriptive statistics and a structural model to	Counties that have one more female urologist per 100,000 residents have	

2014)	urologists relate to rates of bladder cancer death among female patients?	urologists from 2006 and 2009; Florida hospital discharge data from Jan. 2006 - June 2008; Florida Licensure Data; NCI's State Cancer Profiles; Census, BEA, ARF for each market.	explain the distribution of female urologists across counties and the lack of entry.	29.08% fewer female bladder cancer deaths per 100,000 residents. No significant associations between female urologists and male bladder cancer deaths or overall cancer deaths.	
Sabin and Greenwald (AJPH 2012)	What is the association between pediatricians' scores on an implicit bias test (IAT) and racial differences in treatment?	Survey data from 86 academic pediatricians conducted during October and September 2005.	Online survey with IAT tests plus patient vignettes describing children with pain following femur fracture, UTIs, ADHD, asthma.	Pro-white bias in the IAT is significantly correlated with not giving oxycodone to the Black vignette patient in pain after bone surgery ($p<0.05$).	N/A.
Singh and Venkataramani (NBER WP 2022)	How do racial disparities in in-hospital mortality vary with hospital capacity strain?	EHR with time stamps from 2 "highly regarded" academic hospitals serving predominantly Black patients.	OLS with rich controls; Assume that hospital capacity strain at patient arrival is conditionally independent of mortality risk.	No significant differences in conditional patient mortality by race in quintiles 1-4 of hospital capacity strain. At the fifth quintile, Black patients are 0.4 pp more likely to die on a baseline of 2%.	Effects are larger for Black women and Black patients without insurance. Effects driven by high-risk patients.
Wallis et al. (JAMA Surgery 2022)	How does surgeon-patient sex concordance affect post-operative outcomes?	Ontario Health Insurance Plan data; CIHI Discharge Abstracts and Ambulatory Care Reporting Services System; Registered Persons Data; Corporate Provider Database.	Population-based, retrospective cohort study.	Sex discordance was associated with increased likelihood of death (adjusted odds ratio 1.07) and complications (adjusted odds ratio 1.09), but not readmission.	They disaggregate by patient sex and find that effects are driven by male surgeons treating female patients. They also find stronger effects for cardiothoracic surgery.

Appendix Table 2: Effect of Experience and Training on Doctor Skills

Paper	Research Question	Data	Empirical Methods	Results	Heterogeneous Effects?
Chan and Chen (NBER WP, 2023)	How do NPs compare to doctors with respect to patient outcomes and resource use in the ED? How does variation in provider skill vary across and within professions?	Administrative health records from the VHA for ED visits between 01/2017 and 01/2020 (1.1 million cases, 44 EDs) linked to death records.	Use number of NPs on duty as IV for assignment to an NP vs. a doctor on arrival at the ED.	Assignment to an NP increases patient length of stay by 11%, increases cost of care by 7%, and increases 30-day preventable hospitalizations by 20%. Productivity variation is greater within than between each profession.	The NP-physician performance gap is smaller for experienced providers and larger for patients with complex or severe conditions. Many NPs are more skilled than some doctors.
Currie and Zhang (ReStat, 2023)	Are some physicians more effective in promoting patient health? Correlation in effectiveness across domains of patient care? Do effective providers have lower/higher costs?	EHR data from the Veterans Health Administration's Corporate Data Warehouse for 2004 to Feb. 2020, VHA Vital Status files, CDC National Death Index Plus files.	Quasi-random assignment of veterans to PCP teams in the VHA system; value-added measure of provider effectiveness.	PCPs with 1 SD higher mental health effectiveness, circulatory condition effectiveness, or ACSC effectiveness have a 27-44% reduction in adverse outcomes. Effectiveness measures positively correlated. Assignment to a PCP with a 1 SD higher effectiveness reduces mortality 3.6-4.2 % and reduces patient costs 2.5-5.4% over the next three years.	Provider effectiveness increases with provider age and number of patients seen.
Doyle, Ewer, and Wagner (JHE, 2010)	Do residents from highly ranked programs do better than residents from lower ranked programs re: costs and health outcomes?	Veteran's Administration inpatient data 1993-2006; 2000 Census zip code level data.	Residency teams randomly assigned to patients based on the last digit of the SSN.	Patients assigned residents from lower ranked program had 11.96% longer stays and 13.31% higher costs. No differences in health outcomes.	Differences in costs were higher for more serious conditions.
Doyle (NBER WP, 2020)	Does having cardiologists in the ER affect treatment and outcomes for patients with heart failure? Does additional experience with heart failure patients affect outcomes?	Medicare claims data (1998-2002) linked to mortality data; AMA's Masterfile for physician characteristics.	Estimate the effect of the share of physicians of different types in the ER, conditional on hospital*quarter *day-of-week FE.	Controlling for number of physicians available, 1-year mortality falls by 1.10% with each additional cardiologist. Additional cardiologists increase intensity of care. A doctor seeing 10 more heart failure patients yearly reduces mortality 1.2%.	Mortality point estimates larger for patients with higher predicted mortality, in high-volume hospitals, and for patients seen on slow days but differences imprecisely estimated.
Epstein,	Compare effect of initial	Florida and New York	Initial skill defined as	Without hospital FE, initial skill	Privately insured patients

Nicholson, and Asch (AJHE 2016)	skill to the effect of experience in predicting obstetrician performance?	all-payer discharge databases (1992 to 2012); AMA Physician Masterfile; AMA FREIDA identifiers of hospitals with OB residency training.	physician's normalized, risk-adjusted maternal complication rate in the 1st year.	explains much of the variance in performance. After 16 years, it explains 39-75% of performance. With hospital FEs initial skill explains only 1-9%, suggesting better doctors go to better hospitals. Experience explains little.	respond to recent measures of physician skill. Robustness checks with physician "stayers" only show similar results.
Facchini (Health Econ, 2022)	Does the recent volume of C-sections performed affect the outcomes of a surgeon performing a nonelective C-section?	Birth certificates from a large public hospital in Tuscany, Italy (2011 to 2014)	Patients cannot select their surgeon though more skilled surgeons may get harder cases. Include surgeon FEs.	Recent experience defined as #C-sections in the last 4 weeks. A one SD increase in experience reduces NICU admission 13.86% and reduces low APGAR 13.19%.	N/A.
Gowrisankaran, Joiner, and Léger (Management Science 2023)	How are measures of physician practice style and of physician skill correlated in the context of patients visiting the ED?	La Régie de l'as-surance maladie du Québec (RAMQ) data on Montreal patients who visited an ED between April and Dec. 2006.	Identification relies on conditional random assignment of patients within an ED. Physician practice style and skill estimated from physician FEs.	Physicians with more intensive practice style have worse outcomes on average. Practice intensity correlated across conditions, as is skill.	Negative correlation intensive practice style and patient outcomes strongest for appendicitis, weakest for transient ischemic attacks.
Schnell and Currie (AJHE, 2018)	How does a doctor's medical school rank affect their propensity to prescribe opioids? How does this relationship vary over time and between specialties with different levels of training in pain relief?	QuintilesIMS opioid prescription data 2006-2014; US News and World Reports; CMS provider utilization and payment data; ACS data; Mortality data.	FE models (specialty, county of practice, practice address).	Physicians from the lowest ranked medical school are 121% more likely to prescribe any opioids and prescribe 160% more than physicians trained at the top school.	Rank doesn't matter for specialties with pain medicine training. Rank matters less for more recent cohorts. Foreign physicians from low prescribing areas have low prescription rates.
Simeonova, Skipper, and Thingholm (JHR, 2024)	Do health management skills (HMS) of primary care physicians affect medication adherence and hospitalizations for cardiovascular (CV) disease, and CV hospital costs of patients on statins? Do skills change with age?	Danish registry data on population of statin users and their PCPs (01/2004-06/2008). However, cannot observe PCP for 54% of clinics.	Leave-one-out adherence rates for each physician adjusted for patient and physician observables. Event studies after changes in PCP induced by clinic closures or patient moves.	A one SD increase in PCP HMS is associated with a 1.10% increase in medication adherence and 1.47% fall in CV hospitalization. CV hospital expenditures fall by 0.298%. Skill declines with physician age.	N/A.

Van Parys (PLOS One, 2016)	How are variations in ED physicians' treatment of minor injuries related to physician characteristics including experience? Does practice style explain persistence as an ED physician?	All Florida ED visits for minor injuries 2005-2011 matched to Florida Healthcare Practitioner Database; HCUP databases.	OLS assuming little systematic matching of physicians and patients conditional on observables.	Physicians with <2 years of experience spend 4.60% more and perform 3.46% more procedures than physicians with 7+ years. High-cost physicians are 3% less likely to work in a Florida ED 2 years after start.	Differences in care intensity fall with experience after 2-7 years of experience.

Appendix Table 3: Time Pressure and Fatigue

Paper	Research Question	Data	Empirical Methods	Results	Heterogeneous Effects?
Chan (2018) Econometrica	How does ER physician decision- making change over the course of a shift?	Data on physician shifts from the ER in a large, U.S. academic, tertiary-care center 06/2005-12/2012.	Exploits randomness and pre-determination of shifts and overlap in shifts. Counter-factual simulations of patient assignments.	8.70% shorter visits in the 4th to last hour before shift ends, 44.40% shorter in last hour. Patients arriving in last hour have 10.44% more tests/treatments, a 5.7 pp (21.19%) higher likelihood of admission, and 23.12% higher total costs. No significant effects beyond the last hour. No effects found with respect to 30-day mortality or 14-day bounce back.	The effects on workload-adjusted length-of-stay are greater in the daytime and disappear if the index physician has enough time to offload cases to the incoming physician.
Chu et al. (2024) Working Paper	How does cognitive load affect how a physician takes notes, orders tests, and treats patients?	High frequency "click stream" data from EHRs, for patients over 18 at the UCSF ED (2017-2019)	Cognitive load proxied by complexity of patient caseloads. Predict physician orders from past orders; measure deviations in actual orders as a function of load.	When load is high, physicians reduce note editing by 7-14% and increase diagnostic orders by 2-5%, with higher entropy in diagnostic tests. For every 1 SD from expected orders induced by cognitive load, probability of admission increases 3.4 p.p. (14%).	N/A.
Costa-Ramón et al. (JHE 2018)	How does time of delivery affect unscheduled C-sections, and infant health.	6163 births in 4 Spanish public hospitals 2014-2016. Scheduled and breech deliveries excluded.	IV estimation using an indicator for births between 11 p.m. and 4 a.m.	Unplanned C-sections increase by 53.21% between 11 p.m. and 4 a.m. There is a negative effect on 1-minute and 5-minute APGAR (-0.992 and -0.936).	N/A
Freedman et al. (JHE 2021)	Unexpected scheduling changes and decisions of PCPs.	EMR data on all visits to 31 primary care centers in a health system 2005-2015.	Physician FE models with unexpected schedule changes in minutes as the independent variable.	10-minute increase in waiting time reduces total/new (0.19%/0.14%), referrals (0.32%), opioid Rx (0.33%), pap tests (0.39%). Increases scheduled/unscheduled follow ups (0.80%/0.50%), inpatient visits within 14/30 days (1.15%/1.85%), and hospital care within 30 days (0.17%). No effect on ER visits, imaging, antibiotic Rx, diabetes management.	Effects with respect to PT referrals and opioid Rx among opioid-naïve patients are not significant in the baseline specification.
Gruber, Hoe, and Stoye (ReStat 2021)	Studies an English policy limiting ER wait times to 4 hours for 95% of	Records of all visits to public hospitals at the visit level linked	Bunching estimator using the four-hour target. Assumes that only patients around	Wait times fell 8% in patients with wait times of 180-400 minutes, and by 59 minutes for patients moved from the post-threshold period to the pre-threshold period. Increased 30-day	Larger wait time effects and mortality for sicker patients. No significant difference in probability of hospital

	patients at public hospitals.	to vital statistics mortality records for 4/2011-03/2013.	the four-hour mark are affected.	total costs (4.9%); hospital admissions (12.2%); tests in the ER (4.6%); Decreased 30/90-day mortality (13.8%/7.9%); discharge probability (7%); referrals (8.9%). No effect on 1-year mortality, length of stay or number of inpatient procedures.	admission. Most mortality reduction driven by circulatory, respiratory, and digestive problem deaths.
Linder et al. (JAMA IM 2014)	How does time in shift affect the decision to prescribe antibiotics?	Billing and EMRs for visits to 23 Partners HealthCare-affiliated PCPs 05/2011-09/2012.	Logistic regression.	Relative to the first hour of a shift, the adjusted odds ratios of antibiotic prescribing in the 2nd, 3rd, and 4th hours were 1.01 (95% CI, 0.91-1.13), 1.14 (95% CI, 1.02-1.27), and1.26 (95% CI, 1.13-1.41). 44.46% of the sample was prescribed antibiotics.	N/A.
Neprash et al. (JAMA HF 2023)	What is the association between primary care visit length and inappropriate prescribing?	Claims and EHR data from AthenaHealth Inc., 2017.	Descriptive; linear probability models with physician FEs and patient covariates.	An additional minute of visit duration decreases inappropriate antibiotic prescribing 0.11 pp (0.2%), opioid and benzodiazepine co-prescribing for pain 0.01 pp (0.3%), and a prescribing of medications from the Beers List to older adults 0.004 pp (0.4%).	For patients with an anxiety and pain , each additional minute of visit duration decreased dangerous opioid and benzodiazepine co-prescribing 0.05 pp.
Shurtz et al. (RAND, 2022)	Do PCPs increase treatment intensity and screening in response to time pressure caused by absent colleagues?	Administrative data from the largest HMO in Israel covering all primary care visits in Jerusalem 2011-2014.	Event studies at physician-day level. IV for visit length is %caseload missing physicians. (Alt. IV= any doctors missing). Nonparametric methods to bound the ATE.	A 1 minute longer visit increases use of any diagnostic input 4.50% and referrals 7.93%. No significant effects on imaging, pain killer Rx, antibiotic Rx, additional visits.	Effects on use of diagnostic tools bigger for older patients (>60 years) and patients with higher predicted utilization of primary care.
Persson et al. (HE 2019)	How are orthopedic surgeons' decisions affected by the number of patients already seen in a shift?	848 Swedish orthopedic clinic visits spanning 133 work shifts by eight surgeons between 10/2015-12/2015.	Logits with surgeon fixed-effects, assuming patient allocation to time slots is exogenous conditional on observables.	Every additional patient already seen decreases the odds an operation is scheduled by 10.5% (OR = 0.895, CI 0.842 to 0.951). Patients seen in the afternoon are 1.955x more likely to be scheduled for surgery (CI 1.110 to 3.486). Surgery prescribed in 32% of cases.	N/A.
Tai-Seale and McGuire (HE 2012)	Do physicians have a target time per patient?	385 video-taped visits 1998-2000 with 35 PCPs; patient surveys.	Logits on the probability of a topic being the last of the visit.	Topics in the 1st 5 minutes=reference group. Probability of a topic being last increases by 16.8 pp, 26.8 pp, and 35.7 pp for topics raised at 5-10, 10-15, 15+ minutes.	Academic medical centers demonstrated sharpest increase in the shadow price of time.

Appendix Table 4: Peer Effects and Team Dynamics

Paper	Research Question	Data	Empirical Methods	Results	Heterogeneous Effects?
Agha and Molitor (ReStat 2018)	Does proximity to lead investigators in new cancer drug trials increase the propensity to prescribe new drugs?	Medicare Part B claims 1998-2008; Dartmouth Atlas data; FDA drug application data.	DiD, patient location IV (secondary analysis).	Cancer patients in lead investigator's HHR 4.04 pp (36%) more likely to get new cancer drug, with convergence after 4 years. No effect in other authors' HHRs. IV estimates smaller.	Effects bigger in areas with slower drug adoption. Convergence suggests lead investigators are not in areas with higher latent demand for the cancer drug.
Chan (JPE, 2016)	Is doctor shirking (i.e. working slowly to avoid work) reduced when doctors vs. nurse schedulers do patient assignments?	6 years of ED data from an academic medical center. ED had 2 pods of doctors.	Natural experiment in which a nurse-managed pod became doctor-managed, as the other pod was.	The doctor-managed system reduced patient wait times by 13.67% with no significant effects on quality, cost, or utilization.	Patient assignment is more negatively correlated with a physician's number of patients in doctor-managed system (consistent with it being a stronger signal of true workload).
Chan (AEJ: EP, 2021)	How much influence do senior residents have on team decisions? How do junior resident's decisions vary with experience?	Five years of data from the internal medicine residency program of a large teaching hospital.	RE model exploiting discontinuity caused by promotion of junior residents to senior.	There is a jump in the SD of log costs after promotion. Senior residents are responsible for almost all of the variance in decision making within a team of residents.	The jump in practice variation is highest for diagnostic spending (vs. medication, blood work, or nursing). No differences by patient characteristics.
Chen (AER, 2021)	How does the length of time that PCI/CABG surgeons and other hospital physicians have worked together affect patient outcomes?	20% of Medicare claims 2008-2016 linked to Vital Statistics, MD-PPAS 2008–2016, Physician Compare 2014–2017.	1.Restrict to admissions through ED and include FEs for proceduralists. 2.TWFE model with FEs for proceduralists and PCPs.	1 SD increase in shared work experience reduces 30-day mortality by 10 to 14%. Shared work experience decreases use of medical resources and length of stay.	Effect of shared work experience declines with individual physicians' experience, but this decline is small. The effect is larger for more complex cases.
Molitor (AEJ: EP 2018)	How are cardiologists affected when they move to areas with different practice styles?	Medicare fee-for-service claims 1998–2012; AMA Masterfile;	"Movers" design follows cardiologists moves across HRRs; event study and difference-in-	A 1pp increase in cardiac catheterization in the new HRR increases the physician's own rate 0.628 pp (1.36%). A 1pp increase in the rate at the physician's	Effects of moving larger for moves from low to high-intensity areas. Effects similar for moving earlier vs. later in their careers. Effects of moving are larger for

			differences.	hospital leads to a 0.796 pp (1.72%) increase in the physician's own rate.	more marginally appropriate patients.
Silver (ReStud 2021)	How do peer-groups affect speed and outcomes in the ED?	All ED visits from New York (2005-2013). Linked to state physician license register, public physician profiles, and vital statistics mortality data.	Peers vary across shifts. Decompose variation in outcomes attributable to physicians and physician-peer matches. Use peer group as IV for outcomes.	First-Stage: A 10% increase in the speed of a physician's peers increases own speed 1.47% with controls. 2SLS: A peer group that increases a physician's speed by 10% decreases charges by 2.17% with no significant effect on the 30-day mortality of discharged patients.	Physicians work faster in smaller groups and when all of their peers are male. 2SLS: In at-risk patients, peer groups that increase physician speed by 10% decrease charges 2.55% and increase 30-day mortality in discharged patients by 0.2121 pp (5.65%) .

Appendix Table 5: U.S. Financial Incentives

Paper	Research Question	Data	Empirical Methods	Results	Elasticity	Het. Effects?
Papers with Defined Price Elasticities						
Allen, Fichera, and Sutton (HE, 2016)	Examined an English policy that increased payments 24% for outpatient cholecystectomies while inpatient reimbursement were unchanged.	Hospital Episode Statistics from the NHS Information Centre for Health and Social Care from 12/2007-03/2011.	D-in-D using a set of control procedures with similar recommended outpatient rates that were not affected.	Planned outpatient surgeries increased by 27% of baseline mean. Reversion from laparoscopic to open surgery decreased. No effect on deaths or readmissions.	Elasticity of outpatient surgery supply w.r.t. payment: 1.21	N/A.
Alexander and Schnell (AEJ:AE, 2024)	What was the impact of increasing Medicaid PCP payments in 2013 and 2014 to comply with the ACA?	State-level Medicaid reimbursement rates; NHIS (2009–2015); NAEP (2009, 2011, 2013).	D-in-D and event studies exploiting variation in the effect of ACA rule given pre-ACA reimbursement rates.	A $10 rise in payments (a 13.2% rise) decreases prob. doctors decline new Medicaid patients by 0.71pp or 11.5%. Also decreases prob. that parents have trouble finding a doctor for child 25%. Increased payments increased doctor visits, improve reported health, and reduce school absences.	Elasticity of getting and appointment w.r.t. payment: 11.5/13.2 =0.87	Effects on school absences are larger and more precisely estimated for younger students.
Bisgaier and Rhodes (NEJM, 2011)	How does public vs. private insurance affect the probability that specialists will accept new pediatric patients, and wait times?	Experiment with 546 paired calls to 273 specialty clinics. Private insurance pays 60% more.	Audit study. One call with public insurance and one a month later with private insurance.	Private insurance accepted 89.4% of the time, public ins. accepted 34.4% of the time. Medicaid-CHIP callers were 6.2 times more likely to be denied an appointment. Conditional on getting an appointment, Medicaid-CHIP callers waited 22 days longer.	Elasticity of getting an appointment w.r.t. payment: [(89.4-34.4)/34.4]/60 =2.66.	N/A.
Cabral, Carey, and Miller (NBER Working Paper,	How did increased payments to providers of evaluation & management services to dual-eligible beneficiaries under the ACA affect care provision?	20% random sample of Medicare beneficiaries from Master Beneficiary Summary File and medical claims files	DiD and triple differences using non-duals and non-qualifying providers as control groups.	Increased payments increased evaluation & management services for dual-eligible beneficiaries by 6.3% and reduced fraction with no evaluation & management visits by 8.7%.	Elasticity of evaluation & management services/appointments w.r.t. payment: 1.2	Larger effects for younger/white beneficiaries, and beneficiaries not living in HPSAs.

2024)		(2010–2014); Medicaid Analytic Extract (2011–2013)				
Chen and Lakdawalla (JHE, 2019)	Do physician responses to changes in Medicare reimbursement vary with patient income?	Medicare Current Beneficiary Survey (MCBS) 1993- to 2002; Federal Registers from 1993 to 2002.	2SLS: Instruments are changes in fees from 1997 consolidation of Medicare areas and 1999 changes in estimation of expenses.	A 10% increase in patient income increases price elasticity for services 0.051 (53% of the mean). Different physician responses wrt patient income explain 53% of the increase in the gap in services received by high-income vs. low-income patients.	Mean elasticity= 0.095.	0.05 at 10th percentile of patient income. 0.15 at 90th percentile of patient income.
Clemens and Gottlieb (AER, 2014)	How do changes to Medicare physician payment rates affect provision of care, technology adoption, and patient health?	Medicare Part B claims 1993-2005.	Natural experiment: 1997 consolidation of Medicare geographic areas. Event study with nearest-neighbor matching on counties.	Higher fees increase elective procedures and RVUs per physician. Imprecise effects on MRIs by non-radiologists. Increases in hospitalization for AMI within 1 year, but no effect on 4-year mortality. A "1 percent change in reimbursement rates thus translates, on average, into a 2.5 percent change in the physician's net wage."	Elasticities for RVUs per patient w.r.t. payment: Short run =0.82 Medium run =2.01 Long run= 1.46.	Heterogenous effects by patient age and state-level intensity of care. Higher care elasticities for older patients and patients from states with more intense care.
Coudin, Pla, and Samson (HE, 2015)	How did a French reform that increased the proportion of GPs subject to price regulation, affect the provision of health services?	Administrative INSEE-CNAMTS-DGFiP File on physicians for 2005-2008.	Fuzzy RD using increase in the requirements for GPs to "bill freely" in their contracts with public health insurance.	Price regulation increased the supply of medical care by 66.53% and the number of procedures by 84.23%.	Provision of total medical procedures wrt payment= 1.61	Male GPs increase labor supply more and also increase home visits and prescriptions.
Fortin et al. (JAE, 2021)	Compare FFS contracts vs. contracts that pay a per diem plus a smaller amount per service. Effects on care rendered by pediatricians?	Doctor time-use survey linked to records from Health Insurance Organization of Quebec (1996–2002).	Structural discrete choice model with variation from a reform introducing an optional per diem plus payment contract.	Small changes in time spent with patients, but services rendered under mixed remuneration contract decrease by 5-12%.	Elasticity of hours wrt wages ~0. Elasticity of services: -0.124.	Female doctors and younger doctors are more likely to switch to the per diem contract.

Johnson and Rehavi (AEJ:EP, 2016)	How is the probability of C-section affected if the patient is a physician? Is there an interaction with financial incentives?	Confidential CA Vital Statistics data, 1996-2005; CA physician licensure data; TX birth data 1996–2003 and 2005–2007.	Comparison group is educated mothers. Nearest neighbor matching regressions for CA. Hospital fixed effects.	California physicians are 1.17 pp (6.13%) less likely to have an unscheduled C-section at non-HMO hospitals. In Texas physicians are 2.09 pp (6.39%) less likely to receive a C-section. Financial incentives affect C-section rates only among non-physicians.	Elasticity~0 for physician-mothers. Non-zero for other mothers but not computable from paper.	Effects greater for physician parents who specialize in areas related to childbirth.
Papers about Capitation/Managed Care Organizations.						
Dickstein (WP 2017)	Are there differences in how physicians in capitated plans prescribe for depression compare to physicians in non-capitated plans?	MarketScan: 2003-2005 Commercial Claims & Benefit Plan Design Data; County-level IRS Income; National Ambulatory Medical Care Survey.	Structural model, instrumenting drug price with sum of price changes within an insurer's plan for all other drugs.	Prescribers in capitated plans are more likely to choose generic Rx. Patients have higher adherence and less medication switching but also higher relapse rates.	Lower drug switching may promote adherence but has negative effects on patients at highest risk of relapse.	
Ding and Liu (JHE, 2021)	How does capitation affect treatment of lower back pain?	MarketScan Commercial Claims 2003- 2006.	Plan history FEs and physician FEs.	Providers with capitation use 12.2% fewer medical resources to evaluate and treat lower back pain with no effect on relapse probabilities.	Effects are biggest for physical therapy and diagnostic testing. But do capitated providers report all procedures?	
Chorniy, Currie, and Sonchak (JHE, 2018)	How does switching from FFS to MMC affect children's treatment of asthma and ADHD?	60% random sample of all South Carolina (SC) Medicaid enrollees < 17, 2005-2015; Vital Statistics	Staggered roll out of MMC contracts with higher capitated payments for children with chronic conditions; child FEs.	Switching to MMC increased ADHD caseloads by 11.6% and asthma caseloads by 8.2%. No significant effects on hospitalization and increases in ER use.	N/A.	
Physician Detailing						
Agha and Zeltzer (AEJ: EP, 2022)	How do pharma payments affect the prescribing of physicians who only share patients with physicians who receive payments?	Medicare Part D (2014–2016); Open Payments database (2013–2016); CMS Referral Patterns;	Event studies; DiD-style regressions with doctor-drug and drug-quarter-specialty FEs	Peers of physicians who receive payments for speaking, consulting, etc., increase prescribing of the promoted drug 1.8%. Spillovers account for ¼ of increased prescribing	Effects are larger for peer physicians with more shared patients with the physician receiving payments.	

		Physician Compare.		from payments.	
Carey, Daly, and Li (NBER WP, 2024)	How do pharma payments affect the prescribing of physician-administered cancer drugs in Medicare?	Open Payments database; claims from 20% sample of Medicare FFS (2014–2018).	D-in-D and event study models with physician-drug and time-drug FEs.	Payments increase Rx the marketed drug by 4% in the year after payment. No improvement in patient mortality. No elasticity because payment value not reported.	Targeted doctors increase treatment of patients with lower expected mortality.
Carey, Lieber, and Miller (JPubE, 2021)	How does detailing affect physician prescribing behavior in terms of drug efficacy, and use of generics?	20% Medicare Part D 2013-2015; Open Payments database; hand-collected data on drug efficacy.	Event studies with physician by drug FEs	Prescribing of the detailed drug increases by 2.2% in the 6 months following payment. No significant effects on efficacy or transitions to generics.	Results are similar when restricting sample to physicians who receive small payments.
Newham and Valente (JHE, 2024)	How do gifts to doctors from pharmaceutical companies affect antidiabetic drug prescribing patterns and costs?	Open Payments database; Medicare Part D data (2014–2017); demographic and health data from ACS and CDC.	Compare physicians with similar propensities to receive payments and use random timing. Residuals from outcome models regressed on residuals from payment models.	An increase in payments by the average yearly payment of $65 increases Rx of branded antidiabetic drugs by 4.8%, increasing costs of Rx drugs.	Effects are higher for doctors in areas with a higher proportion of patients receiving subsidies for out-of-pocket drug costs for low-income individuals.
Shapiro (MS, 2018)	Compare effect of new information from clinical trials and detailing on PCP prescribing behavior for Seroquel.	AlphaImpactRx monthly panel of 1,762 PCPs 2002-2006 (links self-reported detailing, patient treatment).	Two clinical trials over sample period, plus record of detailing. Examine effects in models with physician and month FEs.	No effect of the clinical trial information. Detailing increased after both trials. Detailing increased Seroquel Rx 26% in the month of the visit.	One third of the increase in prescribing occurred in off-label uses.
Other Papers without Defined Elasticities					
Alexander (JPE, 2020)	When hospitals offer incentives to physicians to lower costs, does it affect (1) who is admitted (2) which hospital they are	New Jersey Uniform Billing Records (2006-2013); AHA annual survey; Medicare cost-to-	D-in-D with doctor FEs using the New Jersey Gainsharing Demonstration as a policy experiment.	The policy doesn't reduce costs or change procedure choice. But lower predicted cost patients are sorted towards participating hospitals.	Effects are less precisely estimated for surgical patients, where there is less opportunity for gaming.

	admitted to, and (3) how intensely they are treated?	charge ratio series.			
Alexander and Currie (EHB, 2017)	What is the effect of private vs. public insurance on propensity to be admitted to hospital from ED? Are effects moderated by capacity constraints?	New Jersey Uniform Billing Records 2006- 2012.	Exogenous variation in hospital bed supply due to local flu conditions; hospital FEs.	In high flu weeks, publicly insured children are .3 p.p. (6.4%) less likely to be admitted for non-flu conditions compared to privately insured children. Outcomes are no worse for marginal children.	Effects are larger when restricting to diagnoses with mid-range admissions rates.
Brekke et al. (JHE, 2019)	How does GP compensation and relationship with patients affect their propensities to issue sick-leave certificates patients need to claim benefits?	Norwegian administrative data 2006–2014 linking health, national insurance, and labor market data.	Physicians see patients both in their own practices and in EDs where they do not face reputational effects. Models with physician and patient FEs.	GPs with a FFS contract are 34.63% more likely to issue sickness certificates for own patients vs. ED patients. For GPs with fixed salaries the gap is 24.15%.	GPs with new practices have similar effects with FFS but not for fixed salary. The effect for fixed salary is driven by relationships with patients. Effects larger in areas with more GPs per capita and where GPs have more openings.
Chernew et al. (JHE, 2021)	How much of the variation in prices for lower-limb MRIs is explained by physician referral patterns vs. patient characteristics?	2013 insurance claims from a large national insurer; data from the company's online price comparison tool; SK&A physician-level dataset.	Restrict to lower-limb MRIs without contrast since these are "shoppable, homogeneous MRI scans." Estimate models with referrer FEs.	Referrer FEs explain 52% of the variance in patient spending on lower-limb MRIs. Patient cost-sharing and characteristics explain less than 1%. Patient HHR FEs explain 2%. Going to the cheapest provider within the same driving distance would reduce spending 35.83%.	The mean vertically- integrated physician refers 52% of patients to a hospital-based MRI provider compared to 19% for non-vertically-integrated physicians.
Clemens et al. (NBER WP, 2024)	How do measures of provider preferences for treatment intensity relate to utilization and spending for commercially insured patients? How do financial incentives mediate these relationships?	Health Care Cost Institute Commercial Claims Database; survey data from Cutler, Skinner, Stern, and Wennberg (2019)	Descriptive analysis following Cutler et al. (2019) with additional covariates to represent different financial incentives in commercial insurance.	Provider preference measures (share Cowboy, Comforter High Follow-Up, Low Follow-Up) are weakly related to utilization and spending, in contrast to Cutler et al. (2019). Private insurance offers lower prices in areas with a higher share of Cowboys/High Follow-Up, offsetting provider preferences.	Relationship between provider preference measures and non-price utilization measures are weaker than relationship between provider preference measures and payments.
Frakes	Does physician behavior	National Hospital	Focus on AMI and C-	After adoption of a national-standard	Disaggregates by whether states

(AER, 2013)	converge towards national averages when states change malpractice laws to consider national rather than local norms?	Discharge Survey (1977-2005), Natality Data (1978-2004); Mortality Data (1977-2004).	section. Event study exploiting variation in states adoption of national-standard rules.	rule, the deviations between state and national C-section rates fall by 4.87 pp (48.31%). Estimates for AMI are noisier. No convergence in outcomes.	have rates that are initially higher or lower rates than the national rate. Convergence occurs in subsamples.
Gupta (AER, 2021)	Effects of the Hospital Readmissions Reduction Program (HRRP) on care quality and admissions for patients with heart attacks, heart failure, and pneumonia?	Medicare fee-for-service claims 07/2006-07/2006; 20% sample of all Medicare beneficiaries.	D-in-D, IV using baseline predicted readmission rate.	HRRP reduced 30-day readmissions by 10.5% and 30-day returns to the hospital by 6.92%. Little effect on admission decisions or upcoding. Increases in procedures for AMI patients and 8.87% fall in 1-year mortality.	Readmission rates lower for patients initially admitted to index hospital, not for those originally seen elsewhere. Government hospitals respond less. Higher volume hospitals and at-risk systems respond more.
Howard and McCarthy (JHE, 2021)	Did a DOJ investigation of Medicare fraud re: implantable cardiac defibrillators (ICDs) change practice?	All-payer data from Florida; ED data from Florida's Agency for Healthcare Administration.	D-in-D using ICD procedures not subject to the investigation as a control.	The investigation plus new checklists that were part of the settlement caused a 22% decline in unnecessary ICD implantations.	The decline in ICDs was stronger for hospitals involved in the lawsuit. Decline for Medicare patients smaller in percent but larger in absolute terms compared to patients with other insurance.
Johnson et al. (NBER 2016)	Are OBs more/less likely to do unscheduled C-sections on own patients? Effects recent patients' laceration rates?	EMR and billing databases for three practice groups.	They use rotating call schedules of OB groups as a plausibly exogenous source of OB assignments.	OBs are 4 pp (25.97%) more likely to perform a C-section and 2.5 pp (25.0%) less likely to use vacuum or forceps on their own patients vs. another OB's.	Higher rates of recent lacerations increase the probability of C-section for an OB's own patients but not for other patients.
Wilding et al. (JHE, 2022)	How did increased stringency of blood pressure targets for patients <80 affect English GPs' treatment and testing decisions for hypertensive patients?	EHRs from Clinical Practice Research Datalink (04/2010-03/2017); Health Survey for England.	D-in-D comparing patients over and under 80; bunching estimators.	Stricter targets did not increase diagnoses of hypertension in new patients but increased antihypertensive Rx 1.2 pp. Doctors did multiple tests when patients failed, reported more patients as exempt from reporting, and increased reports of patients exactly meeting targets.	Lower-performing practices increased reporting of patients as exempt more than higher-performing practices, but other effects were similar. No data on health outcomes.

Note: One could compute detailing elasticities for some of the papers above, but these measures are difficult to interpret because detailing involves more than payment. Carey, Lieber, and Miller (JPubE, 2021) find that effect sizes are very similar when restricting to small payments, suggesting that direct remuneration is not the main reason that detailing affects physician decision making.

Appendix Table 6: Doctor Responses to New Information

Paper	Research Question	Data	Methods	Results	Heterogeneous Effects?
Avdic et al. (JHE, 2024)	New stents were first thought to reduce complications and then to increase them. How did cardiologists respond to new information and guidelines?	Swedish Coronary Angiography and Angioplasty Registry 2002-2011.	Separate models for periods after positive info, after negative info, and after guidelines allow physician-specific intercepts and trends.	Doctors responded more quickly to negative information than to the initial positive information.	Doctors slow to take up new stents were more likely to use the appropriate stent and had better patient outcomes. No heterogeneity within hospitals. Slow responders more likely to practice in teaching hospitals.
Ahomaki, Pitkanen, Soppi, and Saastamoinen (JHE, 2020)	Experiment with letters sent to Finnish doctors who prescribed 100+ paracetamol-codeine pills to a new patient.	National Prescription Register including all purchases, merged to Nordic Product Number and physician characteristics.	D-in-D using new patients where non-targeted physicians are the control. "Treatment" is intent-to-treat.	Significant 6.13 tablet decrease in number of pills purchased by new patients of treated doctors relative to patients of untreated doctors (12.8% of treatment group baseline).	Treatment effects larger for high prescribers. Top 5 specialties have Similar effect size. The decrease in large purchases was greatest in urban areas and not significant in rural areas.
Bradford & Kleit (HE, 2015)	The effect of the 2005 Blackbox warning on NSAID prescriptions, and how it was mediated by advertising, media coverage, and patient characteristics.	EMRs from the Primary Care Practices Research Network; media data from Competitive Media Reporting, Inc. and Lexis/Nexis; NSAID sample dispensation data from IMS health.	Probit models on having active prescription for non-COX-2 inhibitor NSAIDs, COX-2 inhibitor NSAIDs, opioids, and other analgesics.	Blackbox warnings resulted in a 2.8pp (54.90%) decrease in prescriptions for COX-2-inhibitors and 2.8pp (23.14%) increase in prescriptions for a non-COX-2-inhibitor ($p<.001$).	Patients with cardiovascular disease had a similar decrease in prescription of COX-2-inhibitors, but no significant increase in non-COX-2-inhibitors. These patients substituted toward opioids and other analgesics.
Currie and Musen (Working Paper, 2025)	Effect or prior authorization policies on prescribing of antipsychotics to kids on Medicaid.	New hand-collected data on Medicaid prior authorization policies (2005–2020); IQVIA LRx database of psychotropic Rx (2006–2019).	Staggered DiD using state-level rollout of prior authorization policies.	Comprehensive pediatric prior authorization policies reduced prescribing of antipsychotics to children ages 3-5 on Medicaid by 34 –43%.	No spillovers to older children or children on private insurance, suggesting hassle costs instead of information as the primary mechanism behind main findings.
DeCicca, Isabelle, and Malak (HE Letters, 2024)	Effect of Term Breech Trial and its subsequent overturning on C-sections for breech births.	U.S. Birth Certificate Records 1995–2010.	D-in-D using complication-free births as control group.	No effect of original Term Breech Trial on C-section rates. Reversal of trial findings reduced C-sections for breech babies by 15–23%.	Reductions in C-sections greater in counties with younger physicians and more IMGs and among non-white, less educated patients.
Doctor, Nguyen,	Effect of notification of	Opioid dispensing from	RCT with intent-to-	Milligram morphine	N/A

Lev, Lucas, Knight, Zhao, and Menchine (Science, 2018)	patient death by overdose on future opioid prescribing.	California's Prescription Drug Monitoring Program database.	treat analysis. Letters from the Chief Medical Examiner of CA.	equivalents prescribed down 9.7% in treatment vs. control 3 months after intervention.	
Dubois and Tuncel (JHE, 2021)	How did French physicians respond to the 2004 information that SSRIs increase suicidal thinking in children?	Cegedim proprietary longitudinal patient data covering all prescriptions by 386 GPs. Includes doctor and patient demographics, and visit-level information.	D-in-D estimation, older patients are control. Random coefficient discrete choice logit examines choice across drug categories.	Child SSRI prescriptions fell 9.9 pp (19.8%). The baseline effect for adults was -2.8 pp (5.6%). Many physicians decreased prescription of other classes of anti-depressants but substituted to off-label use of other drugs.	25% of the physicians prescribe an SSRI for depression <20% of the time before the warning, and 25% prescribe an SSRI >73% of the time. Over 25% of physicians never prescribe SSRIs to children after the warning.
Howard, David, and Hockenberry (JEMS, 2016)	Variation in surgeon responses to the information that arthroscopic knee surgery is ineffective by whether it is a hospital or a free-standing surgery.	Outpatient claims data from Florida's State Ambulatory Surgery Database, 1998-2000. Surgeons cannot be linked over time. Analysis at facility level.	Triple D-in-D, alternative model using differential trends in the ratio of knee to shoulder surgeries (preferred specification).	Preferred specification: if free-standing centers responded like hospitals the number of surgeries would be reduced 6.27-11.37% on a baseline of 34,000 each year.	Disaggregating by procedure type, the differential decline between free-standing centers and hospital centers is driven by meniscectomies, which have received more insurance company scrutiny.
Howard and Hockenberry (HSR, 2019)	How is physician age related to the response to new information that episiotomies are ineffective?	Pennsylvania Inpatient Hospital Discharge Data (1994–2010)	Descriptive. LPM with hospital FEs.	Physicians who started delivering babies 10 years earlier are 6 pp (19.5%) more likely to perform an episiotomy.	The relationship between physician age and episiotomy rate has decreased over time and is weaker in teaching hospitals, which promote evidence-based medicine.
Kolstad (AER, 2013)	Effects of quality "report cards" for Coronary Artery Bypass Graft (CABG) surgeries. Is provider response profit motivated?	Pennsylvania Health Care Cost Containment Council data for 89,406 CABG surgeries 1994-1995, 2000, and 2002-2003 merged with surgeon tenure. Focus is on the surgeons' mortality rate before report cards less the report card risk-adjusted rate.	Reduced form responses to differences between own mortality rates and other doctors'. Structural model of consumer demand separates "intrinsic" and "extrinsic" motivations.	Counterfactuals indicate that "extrinsic" incentives induced a 3.5% decline in predicted risk-adjusted mortality whereas "intrinsic" incentives induced a 13% decline in predicted risk-adjusted mortality.	The response is larger for surgeons who are worse than other surgeons in their own hospital compared to surgeons who are just worse than expected.

McKibbin (JHE, 2023)	How do physicians change prescribing of off-label cancer drugs in response to new information from RCTs?	Data on FDA approvals and RCT results, 100% Outpatient and 20% Carrier Claims files for Medicare part B, 1999-2013.	Event studies comparing drug-cancer pairs with and without newly presented RCT evidence from academic conferences.	8 quarters after a conference, prescriptions of drugs with confirmed efficacy up 192%. Prescribing falls by 33% over 8 quarters with negative information.	Responses discontinuous around p-value 0.05. When the abstract describing the RCT has no mention of improvements in quality of life or side effects, adoption and de-adoption rates are less asymmetric.
Olson and Yin (HE, 2021)	Physician responses to changes in drug labeling from the FDA's 1997 Pediatric Exclusivity provision (provides 6 months of exclusivity in return for conducting Pediatric trials).	Prescription data from NAMC; Label changes and exclusivity from FDA; journal publication data from Benjamin et al. (2006) and PubMed; IMS health data on drug promotions; disease prevalence from MEPS.	D-in-D with treatment group defined as children <18 years old and controls as adults >35 (using a zero-inflated negative binomial model).	In their preferred specification, the marginal effect of a pediatric label change is 2.09 fewer prescriptions (12.67 %) for children.	Negative information added to the label reduces prescribing more than positive information. Magnitudes are larger for physicians in solo practice. No clear pattern by child age group. Estimates somewhat sensitive to included controls.
Persson et al. (NBER WP, 2021)	Do doctors consider the diagnosis of an older sibling when evaluating children for ADHD?	Swedish population register 1990-2018, (2016 for HS records); prescription drug claims July 2005-Dec. 2017; birth records data from NHBW, 1996-2016.	Birthday cut-off RD using older sib or cousin's birth date and school eligibility cutoffs to use "young for grade" sib's higher prob. of ADHD diagnosis.	An older sibling born after the school entry cutoff decreases the probability of ADHD diagnosis by 0.59 pp (12.04%) and decreases the probability of ADHD drug claims by 0.55 pp (9.82%). Smaller results for cousins.	Effects on younger siblings are greater before older siblings graduate from HS. Spillovers greater in cities with more funding for special needs children. Cousin spillover effects are greater when cousins are in the same municipality.
Sacarny, Yokum, Finkelstein, and Agrawal (HA 2016)	Effect of letters from Medicare to outlier prescribers of controlled substances on future opioid prescriptions.	CMS Integrated Data Repository-- records for prescription drugs covered by Medicare Part D with prescriber ID.	RCT with analysis of intent-to-treat.	Statistically insignificant increase of 0.8% relative to the control mean after 90 days, 95% CI (-1.38%, 2.91%).	No evidence of heterogeneity by prescriber specialty, geographic region, prescribing pre-treatment, and whether the physician had been investigated for fraud.
Sacarny, Barnett, Le, Tetkoski, Yokum, and Agrawal (JAMA Psych, 2018).	Effect of three letters sent by Medicare to outlier prescribers of quetiapine on future quetiapine prescriptions.	100% Medicare claims data 2013-2017; enrollment data 2015-2017; risk-adjustment data 2013-2014.	RCT with analysis of intent-to-treat.	11.1% fewer days over 9 months vs. control mean (11.99% of the sample mean). Effects lasted 2+ years. No negative effects on patients.	The reduction in prescribing was larger for patients with low-value indications and smaller for guideline-concordant patients.

Wu and David (JHE, 2022)	How did relative procedural skill affect the prob. that doctors abandoned laparoscopic hysterectomy after a negative info shock about the safety of the procedure?	All hospital inpatient and outpatient visit data for patients receiving hysterectomies in Florida (January 2012 – Sept. 2015).	Leave-one-out IV for physician skill at laparotomy/ laparoscopic hysterectomy; DiD event study estimates before/after 2014 FDA announcement.	A 1 SD increased in relative skill in laparoscopic hysterectomy decreased prob. of abandoning the procedure by 4.6–4.9 p.p. (6.2–6.5% reduction from pre-period mean). Only top laparotomy doctors increased laparotomies.	Patients with characteristics that indicate less appropriateness for the laparoscopic procedure had greater reductions in likelihood of receiving a laparoscopic procedure after the announcement.

Appendix Table 7: Heuristics and Guidelines

Paper	Research Question	Data	Empirical Methods	Results	Heterogeneous Effects?
Abaluck et al. (NBER WP 2021)	How does the proportion of physicians following guidelines for anticoagulants for atrial fibrillation patients change after 2006 guidelines? Is lack of implementation due to awareness or nonadherence?	Text mining of EMRs from the VA for patients newly diagnosed with atrial fibrillation between Oct. 2002-Dec. 2013; Patient-level data for 8 clinical trials of anticoagulants.	Causal-forest model to estimate heterogenous treatment effects using data from eight RCTs; Chernozhukov et al. (2018) approach to calculating best linear predictions of conditional average treatment effects.	After 1st mention of guidelines, physicians become more compliant. Stricter adherence could prevent 24% more strokes.	Most departures from guidelines are not justified by measurable treatment effect heterogeneity (though RCTs were not originally randomized on the observables analyzed).
Almond et al. (QJE, 2010)	Does the care of newborns change discretely at the threshold for being classified "very low birthweight" and does this affect mortality?	NCHS linked birth/infant death files (1983-1991 and 1995-2002); linked birth, death, hospital discharge data from California (1991 -2002); HCUP for AZ, NJ, MD, NY.	RD centered around threshold of 1,500 grams.	Relative to the means just above the threshold, VLBW classification has an 11.11% effect on spending and a 5.93% effect on length of hospital stay.	Effects are greater for non-NICU and Level 0/1/2 NICU hospitals than for Level 3A-3D NICU hospitals.
Coussens (Working Paper 2022)	Do doctors use simple heuristics in patient age to make treatment decisions for ischemic heart disease (IHD)?	Truven Commercial Claims and Encounters database 2005-2013; ED records from a large Boston-area hospital 01/2010-05/2015.	Regression discontinuity centered at age 40	Turning 40 increases the probability of being tested, diagnosed, or admitted for IHD by 0.887pp, 0.131pp, and 0.068pp, respectively. Relative changes compared to intercepts are 9.51%, 19.29%, and 17.80%, respectively.	Effects are larger for women and patients presenting without chest pain. Effects are also stronger when the ED is less busy and in the 1st half of a physician's shift.
Cuddy and Currie (PNAS, 2020)	What is the probability that adolescents with private insurance receive appropriate care following an initial diagnosis of mental illness? What factors are related to the type of care received?	Claims data for a large national insurer. Children covered for at least a year between 2012 and 2018 who were ever diagnosed with a mental health condition.	Observational study using linear probability models. Define "red-flag" treatment as prescribing that falls outside accepted guidelines.	Only 75% of adolescents receive follow-up care within 3 months. Of those receiving drugs, 44.85% receive "red flag" drugs. Composition of clinicians affects treatment: More psychiatrists→more drug use vs. more therapists →more therapy.	Any treatment, drug treatment, red-flag drugs increase with age. Girls more likely to be treated, to get therapy, and to get be red-flag drugs. Variation *across* zip codes explains less than half of overall treatment variation.
Cuddy and Currie (JPE, forthcoming)	Would adherence to guidelines improve outcomes? Is there a	Claims data for a large national insurer. Children diagnosed with depression	Instrument individual prescriptions with area-level practice style	Outcomes for red-flag vs. grey-area vs. FDA approved drug treatment after 24	P(drug treatment) is higher for girls, older children, and children whose 1st visit

	difference between "grey-area" prescribing sanctioned by professional societies but not by FDA, and "red-flag" prescribing not sanctioned by either?	or anxiety for the first time 2012-2018. Measures of local practice style computed from IQVIA and from the claims data.	measures interacted with patient characteristics (use Lasso to choose instrument set).	months: P(self-harm): 5.8%; 4.9%; 3.8%. P(ED or hosp.): 33.6%; 18.6%; 26.8%. Total costs: $9557; $1745; $9658. Red-flag has highest costs and worst outcomes.	resulted in hospitalization.
Currie and MacLeod (Econometrica 2020)	Would adherence to professional guidelines improve outcomes? Does the answer to this question vary with the physician's skill?	Claims data for a large national insurer. Adults ever diagnosed with depression 2013-2016; NPPES; Experimental propensity is measured using prescription dispersion across drugs in IQVIA Xponent prescription data base.	Patient FE models of effects of having more experimental doctors and of violations of guidelines. Simulations measure benefits of experimentation for different skill groups. (Psychiatrists assumed more skilled than GPs).	Violations of professional guidelines are associated with worse subsequent outcomes (spending, hospitalizations, ED visits) for all patients.	Among patients seeing psychiatrists, switching to a more experimental doctor improves outcomes (a 0.25 increase reduces P(ED visit or hospitalization) by 10.2%). No effect of experimentation with less skilled doctors.
Geiger et al. (JAMA HF, 2021)	What is the effect of a designation of "advanced maternal age" (AMA) on prenatal care and birth outcomes?	Claims and monthly enrollment data from a large, nationwide commercial insurer 2008-2009; zip-code level public ACS data.	Focus on discontinuities in care for mothers 35+ on expected delivery date. Donut RD excluding women with due dates within 7 days of their 35^{th} birthday.	AMA increases screening, specialty visits; decreases perinatal mortality by 0.39pp or 42.39% of sample mean. No effects on severe maternal morbidity, preterm birth, or low birth weight.	As a percentage of baseline the effects on prenatal care services and perinatal mortality are much greater for low-risk pregnancies than for the full sample.
Kowalski (ReStud, 2023)	Are women who are more likely to receive mammograms different from women who are less likely? How does the probability of being "over-diagnosed" vary with the propensity to receive mammograms?	RCT data from the Canadian National Breast Cancer Screening Study (CNBSS) linked to cancer registries and the mortality data. Allows long-term follow up to see cancers that are detected but would not have caused symptoms.	Extension of Imbens and Angrist (1994) framework in the context of an RCT (which provides identifying variation).	In women who are treated compliers w.r.t. screening guidelines, 14% of breast cancers are "over-diagnosed". For always takers, over 36% of breast cancers are over-diagnosed. Results suggest current guidelines should be revised to reduce mammography.	Women who are more likely to receive mammograms are healthier and of higher socioeconomic status on average.
Ly (Annals of Emergency Medicine, 2021)	Are physicians more likely to test for pulmonary embolism (PE) in the ED when they recently treated a patient with PE?	National EHR data from the VA Corporate Data Warehouse (2011–2018)	Linear probability model with time and physician FEs and clinical and demographic covariates	In the first 10 days after treating a patient with PE, physicians increase testing for PE by 15%. No change in testing behavior in the 50 days after the first 10 days.	N/A.

Ly, Shekelle, and Song (JAMA Internal Medicine, 2023)	Do physicians delay testing for pulmonary embolism (PE) in patients with congestive heart failure presenting in the ED with shortness of breath when congestive heart failure is documented in triage?	National EHR data from the VA Corporate Data Warehouse (2011–2018)	Linear probability model with time and physician FEs and clinical and demographic covariates	The mention of congestive heart failure in triage reduced testing in the ED by 4.6 p.p. (34.8%) and delayed testing in the ED by 15.5 minutes (20.5% increase). Patients were 0.15 p.p. (65.2%) less likely to be diagnosed with PE in the ED but no difference in diagnosis of PE w/in 30 days.	N/A.
Olenski et al. (NEJM, 2020)	Do physicians use simple heuristics in patient age to make treatment decisions for Coronary Artery Bypass Graft Surgery (CABG)?	Medicare data from 2006 to 2012.	Regression discontinuity at age 80.	Patients admitted in the 2 weeks after their 80th birthday were 1.7pp (28.05%) less likely to get CABG than patients admitted 2 weeks before their birthday.	N/A.
Singh (Science, 2021)	Do physicians switch delivery mode after a complication with their previous patient?	EHR (2000–2020) from the obstetric wards of two academic hospitals.	Linear probability model with time, physician, and hospital FEs and clinical and demographic covariates	After a complication with a C-section, physicians are 3.4% more likely to use a vaginal delivery with the next patient. After a complication with a vaginal delivery, physicians are 3.6% more likely to use a C-section with the next patient.	Effects are larger for more experienced physicians.

Appendix Table 8: Technology

Paper	Research Question	Data	Empirical Methods	Results	Heterogeneous Effects?
Agarwal et al. (NBER WP 2024)	How do radiologists use AI predictions and clinical histories in diagnosis? What is optimal use of AI?	Patient cases from Stanford University healthcare; data from an experiment on radiologist decisions and decision time.	2x2 experiment with radiologists. Add AI prediction, clinical history from referring doctor, or both; random forest regression.	AI does not improve performance. Access to clinical history reduces deviation from diagnostic standards by 4%. Optimal to have AI decide cases when confident and radiologists decide all other cases w/o AI.	When the AI tool has high confidence, AI improves radiologist diagnosis. When the tool has low confidence, AI worsens radiologist diagnostic accuracy.
Agha (JHE 2014)	Impact of EMRs plus clinical decision supports on quality and cost of care.	20% sample of Medicare claims, 1998- 2005; Health Information and Management System Survey.	Exploits differential timing of Health Information Technology (HIT) adoption at hospital level w FE.	HIT adoption increases spending 1.3%. No effect on 1-year patient mortality, length of stay, #physicians seen within a year of admission, intensity of care, 30-day readmissions, complications, or an index of care quality.	No evidence of higher returns to more comprehensive HIT systems. Do not see larger effects in larger hospitals.
Alpert, Dystra, and Jacobson (AEJ:EP, 2024)	How much does information versus hassle costs from MA-PDMPs affect opioid prescribing?	Claims data from Optum's Clinformatics Data Mart (2006–2016).	DiD and event studies using policy change in Kentucky. Triple differences comparing opioid naïve and non-naïve patients.	Hassle and information explain 69% and 31% of fall in opioid Rx respectively. MA-PDMPs reduce opioid Rx 6.8% for opioid naïve patients, 10.6% for non-naïve patients, and 16% for patients with opioid-inappropriate conditions.	Declines in prescribing to opioid non-naïve patients occur for patients with history of doctor shopping or high dose/quantity of opioid use.
Arrow, Bilir, and Sorenson (AEJ: AE 2020)	Does access to an electronic database for pharmaceuticals affect doctors' prescribing of cholesterol drugs?	IMS Health Xponent database 2000-2010; data from the firm that owns the studied electronic reference database.	Models with zip-code-month FEs, physician FEs, and physician-specific time trend; IV doctor's access using share of area doctors using database.	Database increases prescribing of generic Rx in its 1st year by 1.3 pp (3.7%). No effect on new branded Rx. New and old generic Rx increase; Old branded Rx decrease. Providers prescribe 0.7% more unique Rx.	In zip codes with more pharmaceutical patenting, database has less effect on drug adoption. Effects stronger for providers who access the database more frequently upon adoption.
Buchmueller and Carey (AEJ: Economic Policy, 2018)	How do MA-PDMPs versus PDMPS without must-access provisions affect opioid use in Medicare?	PDMP info from Prescription Drug Abuse Policy System; 5% Medicare beneficiaries in Part D and FFS in any year 2007–2013.	DiD and event study models using variation in state-level policy.	Without must-access provisions PDMPs have no effect on opioid utilization. MA-PDMPs reduce doctor shopping by 8% and pharmacy shopping by 15%. Neither PDMP significantly affects opioid poisoning rates.	Effect sizes are larger must access provisions are broader.

Buchmueller, Carey, and Meille (Health Economics 2020)	Effect of Kentucky's must-access PDMP program on opioid prescribing.	Kentucky (2006-2016) and Indiana (2012-2016) PDMPs; CDC data on opioid prescriptions; ARCOS 2006-2016.	DiD comparing Kentucky (treated) to Indiana (control).	Quarterly morphine equivalents per capita fell 11–13% in KY vs. IN. Providers prescribing any opioids fell by 3.8 pp (5%). The number of patients prescribed fell 16% among providers prescribing any opioids.	Providers who initially prescribed fewer opioids were more likely to stop prescribing. Reductions in prescribing greater for patients who used opioids multiple times and doctor-shoppers.
Dahlstrand (Working Paper, 2021 updated 2024)	How much could patient outcomes be improved by using an algorithm to match patients and GPs?	Data from Sweden's largest digital healthcare platform (2016–2018) matched to Swedish registry health data.	Physician skill estimated using leave-one-out measures with shrinkage. Match effects exploit the platform's conditional random assignment of patients.	Using an algorithm with positive assortative matching could reduce avoidable hospitalizations by 8%, all hospitalizations by 3%, and counter-guideline antibiotic Rx by 3%.	Effects are smaller for patients seeing a doctor within the day/hour. In urban areas, similar improvements are possible by restricting matches to doctors patients can travel to see in person.
Ellyson, Grooms, and Ortega (Health Economics 2022)	Do the effects of must-access PDMPs vary by specialty?	CMS Part D public use files 2010–2017; AMA Physician Masterfile; PDMP start dates from Prescription Drug Abuse Policy System.	DiD and event study.	Primary care doctors decrease opioid prescribing by 4% after MA-PDMP implementation. No significant effect for providers in IM, EM, surgery, palliative care, oncology, and pain medicine.	Primary care and IM providers with initially low prescribing stop prescribing opioids after MA-PDMP.
Goetz (International Journal of Industrial Organization 2023)	How does an increase in competition on a telehealth platform for talk therapy affect providers' pricing and exit decisions?	Therapist data collected from Psychology Today in 2020; controls from Canadian government sources and Facebook's Movement Range maps.	Propensity score matched DiD exploiting change in how platform shows providers to patients. For areas with <20 providers, platform made providers outside area visible.	Increased competition caused by the platform displaying more providers decreases the likelihood that affected providers provide sliding scale discounts by 8.9%.	Providers with more training respond to competition by stopping sliding scale offers; providers with less training exit the platform. Bigger effects on late adopters of teletherapy.
Horwitz et al. (NBER Working Paper 2024)	How do Certificate of Need (CON) laws affect imaging? How does this vary by the value of imaging?	Hand-coded laws; AHA's Annual Survey of Hospitals 2018; accreditor data on free-standing CT/MRIs; 20% sample Medicare FFS claims 2009–2014.	RDD at state borders where one state has a CON law and the other does not.	The prob. of receiving an MRI is 2% lower on the CON side of the state border, compared to the mean on the non-CON side. Overall, no effect on prob. of a CT.	The prob. of receiving a high-value MRI does not change at border, the prob. of receiving a high-value CT on the CON side falls by 6% of non-CON mean. Low-value imaging falls 20–26%.

McCullough et al. (Health Affairs 2010)	How is quality of care related to EMR adoption 2004-2007?	AHA's annual survey; Health Information and Management Systems Society Analytics database.	OLS with hospital and year fixed effects, coefficient of interest is on the one-year lag of EMR adoption.	Pneumococcal vaccination rates up 2.1pp (3.2%); use most appropriate antibiotic for pneumonia up 1.3pp (1.6%). No effect on other quality of care measures studied.	The relationship between quality measures and EMR adoption is stronger in academic vs. non-academic hospitals.
Miller and Tucker (JPE 2011)	Does EMR adoption lower neonatal mortality.	Linked birth and infant death data 1995–2006; AHA surveys; BEA Regional Accounts; CBP; HIMS Analytics Data; Georgetown Health Privacy Project; Lexis-Nexis.	Construct balanced county-level panel over 12 years. OLS w county and year FEs; IV for EMR adoption using state medical privacy laws.	A 10% increase in EMR adoption reduces neonatal mortality by 3%. Reductions are due to prematurity and complications not to accidents, SIDS, or congenital defects.	Larger effects when EMRs combined with digital storage, and obstetric-specific/decision support technologies. Larger gains for mothers who are Black, Hispanic, unmarried, or have < high school education.
Neumark and Savych (American Journal of Health Economics, 2023)	How do MA-PDMPs and laws that limit initial opioid Rx length for patients with work-related injuries?	Workers Compensation Research Institute claims for workers injured Oct. 2009 – March 2018.	DiD using state-level variation in laws.	Laws that limit opioid Rx length have no effect on opioid Rx (w/pre-trend w/o state trends). MA-PDMPs reduce opioid Rx on intensive but not extensive margin. For neuro spine pain, non-opioid pain Rx increase 14%.	Effects of MA-PDMPs are larger for neurologic spine pain, spine sprains and strains, and other sprains and strains cases.
Obermeyer et al. (Science 2019)	Is there racial bias in algorithms used to target care for high-risk patients? Do doctors correct for algorithmic biases?	Data from all primary care patients enrolled in risk-based contracts at a large academic medical center, 2013-2015.	Descriptive statistics and simulations.	Conditional on chronic condition, Black patients get less recommended care. Black patients have 26% more chronic conditions at the 97th percentile of the risk score. Simulations suggest that physicians do not counteract bias in the algorithms.	Algorithm was trained on spending. Conditional on diagnosis, Black patients have lower spending and algorithm reproduces this bias. Changing algorithm to target health outcomes could potentially resolve the problem.
Mullainathan and Obermeyer (QJE 2022)	Ask how the actual decision to test for heart attacks differs from algorithmically predicted risks and explore health implications.	"Large urban hospital's" HER from Jan. 2010 to May 2015 linked to Social Security Death Index; 20% sample Medicare FFS claims Jan. 2009 to June 2013.	Descriptive comparisons of output from risk model and actual physician decisions; shift-to-shift variation in average testing rates associated with triage team.	Physicians over test low-risk patients and under test high-risk patients because they focus on salient and representative symptoms, ignoring more complicated predictors of risk. High risk patients who arrive at the ED during high-testing shifts have 32% lower 1-year mortality.	Stress testing is more overused than catheterization. More experienced physicians test less but more accurately target tests toward high-risk patients.
Sacks et al. (JHE,	What are the	Commercial claims	DiD using state-level	MA-PDMPs decrease hazard of a	Increases in new opioid Rx in

2021)	effects of MA-PDMPs and laws that limit initial opioid Rx length on opioid-naïve patients?	from "large, national insurer" (20% sample and 100% sample for patients w/opioid Rx) Jan. 2007–Apr. 2018.	variation in laws.	new opioid Rx by 4.7%. Laws that limit initial Rx length increase hazard of new opioid Rx by 8.7%—reductions in Rx for >7 days are more than offset by increase in Rx for <7 days.	response to laws that limit initial opioid Rx length are stronger for PCPs, providing evidence that these laws may inadvertently signal that short prescriptions are safe.
Van Parys and Brown (NBER WP 2023)	Did broadband access improve the outcome of joint replacement outcomes?	Federal Communication Commission data on broadband roll-out; Medicare Current Beneficiary Survey; TM Claims 1999–2014.	DiD exploiting staggered rollout of broadband; discrete choice model	Broadband access explains 16% of the improvement in joint replacement outcomes between 1999-2008. 10% stems from patients seeking better providers and 6% stems from improvements in care conditional on patient demand.	Improvements in outcomes due to hospital access to broadband are driven by hospitals in markets with less competition.
Zeltzer et al. (JHE 2023)	How does the adoption of a digital device to assist with telehealth visits affect health care?	EHR data from Isreali Clait Health Services (an HMO covering ~half the Israeli population) from 2018–2022.	Matched DiD and event study.	Device-assisted telemedicine increases primary care visits 12%, increases antibiotic use 15.6%, and decreases urgent care/ED/inpatient visits 11–24% compared to baseline mean.	Adults have a smaller increase in primary care use and a larger decrease in urgent care/ER/impatient visits than pediatric patients.
Zeltzer et al. (JEEA 2024)	Impact of increased access to telemedicine during COVID-19 after lockdowns lifted were in May–June 2020.	EHR data from Israeli Clait Health Services from January 2019 to June 2020.	DiD at the patient level. Treatment is a patients' physicians' propensity to use telemedicine during the initial March–May 2020 lockdown.	Having a PCP who was a high user of telemedicine increased the prob. of a primary care visit by 3.6% but reduced visit costs by 5.7% (of the pre-lockdown mean). Visits had fewer Rx and referrals. No evidence of more missed diagnoses for patients of high adopters.	Effects measured in % changes with respect to baseline are similar across patient age, gender, and SES. Reduction in Rx larger for providers who prescribed more in the pre-period.

Glossary of Table Terms

AHA – American Hospital Association
AKM– Abowd, Kramarz, and Margolis (1999)
AMA – American Medical Association
AMI/MI –Acute myocardial infarction
ATE—Average Treatment Effect
CCI—Charlson Comorbidity Index
CDC – Center for Disease Control and Prevention
CMS –Centers for Medicare and Medicaid Services
CPOE – Computerized provider order entry
DEA – Drug Enforcement Authority
D-in-D – Difference in differences
DO – Doctor of Osteopathic Medicine
ED/ER – emergency department
EMR/EHR – Electronic medical/health record
FDA— Food and Drug Administration (United States)
FE – Fixed effects
FFS—fee-for-service
GP—General Practitioner
HCUP – Health care utilization project
HIT – Health information technology
HRR – Hospital referral regions (from the Dartmouth Atlas)
IV –Instrumental variable
MA-PDMP – Must-Access Prescription Drug Monitoring Program
MD – Medical Doctor
MMC—Medicaid managed care
NCHS -- National Center for Health Statistics
NHS—National Health Service (U.K., Norway)
NPI – National Provider Identifier
OR – Odds ratio
PCP –Primary care provider
PDMP – Prescription drug monitoring program
pp – percentage point
PSI – Patient safety indicator
RCT – Randomized controlled trial

RD—Regression discontinuity
Rx—Prescription
SES – Socioeconomic status
SSRI—Selective Serotonin Reuptake Inhibitor
VHA—Veterans Health Administration (United States)

Bibliography for Table Only Citations (see main text for others)

Abaluck, Jason, Leila Agha, Jr Chan David C., Daniel Singer, and Diana Zhu. 2021. "Fixing Misallocation with Guidelines: Awareness vs. Adherence." Working Paper 27467. National Bureau of Economic Research. https://www.nber.org/papers/w27467.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2024. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." Working Paper 31422. National Bureau of Economic Research. https://doi.org/10.3386/w31422.

Agha, Leila. 2014. "The Effects of Health Information Technology on the Costs and Quality of Medical Care." *Journal of Health Economics* 34 (March):19–30. https://doi.org/10.1016/j.jhealeco.2013.12.005.

Agha, Leila, and David Molitor. 2018. "The Local Influence of Pioneer Investigators On Technology Adoption: Evidence From New Cancer Drugs." *The Review of Economics and Statistics* 100 (1): 29–44. https://doi.org/10.1162/REST_a_00670.

Agha, Leila, and Dan Zeltzer. 2022. "Drug Diffusion through Peer Networks: The Influence of Industry Payments." *American Economic Journal: Economic Policy* 14 (2): 1–33. https://doi.org/10.1257/pol.20200044.

Ahomäki, Iiro, Visa Pitkänen, Aarni Soppi, and Leena Saastamoinen. 2020. "Impact of a Physician-Targeted Letter on Opioid Prescribing." *Journal of Health Economics* 72 (July):1–22. https://doi.org/10.1016/j.jhealeco.2020.102344.

Alexander, Diane. 2020. "How Do Doctors Respond to Incentives? Unintended Consequences of Paying Doctors to Reduce Costs." *Journal of Political Economy* 128 (11): 4046–96. https://doi.org/10.1086/710334.

Alexander, Diane, and Janet Currie. 2017. "Are Publicly Insured Children Less Likely to Be Admitted to Hospital than the Privately Insured (and Does It Matter)?" *Economics & Human Biology*, In Honor of Nobel Laureate Angus Deaton: Health Economics in Developed and Developing Countries, 25 (May):33–51. https://doi.org/10.1016/j.ehb.2016.10.005.

Alexander, Diane, and Molly Schnell. 2024. "The Impacts of Physician Payments on Patient Access, Use, and Health." *American Economic Journal: Applied Economics*. https://doi.org/10.1257/app.20210227.

Allen, Thomas, Eleonora Fichera, and Matt Sutton. 2016. "Can Payers Use Prices to Improve Quality? Evidence from English Hospitals." *Health Economics* 25 (1): 56–70. https://doi.org/10.1002/hec.3121.

Almond, Douglas, Joseph J. Doyle Jr., Amanda E. Kowalski, and Heidi Williams. 2010. "Estimating Marginal Returns to Medical Care: Evidence from At-Risk Newborns." *The Quarterly Journal of Economics* 125 (2): 591–634. https://doi.org/10.1162/qjec.2010.125.2.591.

Alpert, Abby, Sarah Dykstra, and Mireille Jacobson. 2024. "Hassle Costs versus Information: How Do Prescription Drug Monitoring Programs Reduce Opioid Prescribing?" *American Economic Journal: Economic Policy* 16 (1): 87–123. https://doi.org/10.1257/pol.20200579.

Alsan, Marcella, Owen Garrick, and Grant Graziani. 2019. "Does Diversity Matter for Health? Experimental Evidence from Oakland." *American Economic Review* 109 (12): 4071–4111. https://doi.org/10.1257/aer.20181446.

Angerer, Silvia, Christian Waibel, and Harald Stummer. 2019. "Discrimination in Health Care: A Field Experiment on the Impact of Patients' Socioeconomic Status on Access to Care." *American Journal of Health Economics* 5 (4): 407–27. https://doi.org/10.1162/ajhe_a_00124.

Arrow, Kenneth J., L. Kamran Bilir, and Alan Sorensen. 2020. "The Impact of Information Technology on the Diffusion of New Pharmaceuticals." *American Economic Journal: Applied Economics* 12 (3): 1–39. https://doi.org/10.1257/app.20170647.

Avdic, Daniel, Stephanie von Hinke, Bo Lagerqvist, Carol Propper, and Johan Vikström. 2024. "Do Responses to News Matter? Evidence from Interventional Cardiology." *Journal of Health Economics* 94 (March). https://doi.org/10.1016/j.jhealeco.2023.102846.

Bisgaier, Joanna, and Karin V. Rhodes. 2011. "Auditing Access to Specialty Care for Children with Public Insurance." *New England Journal of Medicine* 364 (24): 2324–33. https://doi.org/10.1056/NEJMsa1013285.

Bradford, W. David, and Andrew N. Kleit. 2015. "Impact of FDA Actions, DTCA, and Public Information on the Market for Pain Medication." *Health Economics* 24 (7): 859–75. https://doi.org/10.1002/hec.3067.

Brekke, Kurt R., Tor Helge Holmås, Karin Monstad, and Odd Rune Straume. 2018. "Socio-Economic Status and Physicians' Treatment Decisions." *Health Economics* 27 (3): e77–89. https://doi.org/10.1002/hec.3621.

———. 2019. "Competition and Physician Behaviour: Does the Competitive Environment Affect the Propensity to Issue Sickness Certificates?" *Journal of Health Economics* 66 (July):117–35. https://doi.org/10.1016/j.jhealeco.2019.05.007.

Buchmueller, Thomas C., and Colleen Carey. 2018. "The Effect of Prescription Drug Monitoring Programs on Opioid Utilization in Medicare." *American Economic Journal: Economic Policy* 10 (1): 77–112. https://doi.org/10.1257/pol.20160094.

Buchmueller, Thomas C., Colleen M. Carey, and Giacomo Meille. 2020. "How Well Do Doctors Know Their Patients? Evidence from a Mandatory Access Prescription Drug Monitoring Program." *Health Economics* 29 (9): 957–74. https://doi.org/10.1002/hec.4020.

Button, Patrick, Eva Dils, Benjamin Harrell, Luca Fumarco, and David Schwegman. 2020. "Gender Identity, Race, and Ethnicity Discrimination in Access to Mental Health Care: Preliminary Evidence from a Multi-Wave Audit Field Experiment." Working Paper 28164. National Bureau of Economic Research. https://doi.org/10.3386/w28164.

Cabral, Marika, Colleen Carey, and Sarah Miller. 2021. "The Impact of Provider Payments on Health Care Utilization of Low-Income Individuals: Evidence from Medicare and Medicaid." Working Paper. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w29471.

Cabral, Marika, and Marcus Dillender. 2024. "Gender Differences in Medical Evaluations: Evidence from Randomly Assigned Doctors." *American Economic Review* 114 (2): 462–99. https://doi.org/10.1257/aer.20220349.

Carey, Colleen, Michael Daly, and Jing Li. 2024. "Nothing for Something: Marketing Cancer Drugs to Physicians Increases Prescribing Without Improving Mortality." Working Paper. National Bureau of Economic Research. https://doi.org/10.3386/w32336.

Carey, Colleen, Ethan M. J. Lieber, and Sarah Miller. 2021. "Drug Firms' Payments and Physicians' Prescribing Behavior in Medicare Part D." *Journal of Public Economics* 197 (May):104402. https://doi.org/10.1016/j.jpubeco.2021.104402.

Chan, David C. 2016. "Teamwork and Moral Hazard: Evidence from the Emergency Department." *Journal of Political Economy* 124 (3): 734–70. https://doi.org/10.1086/685910.

———. 2018. "The Efficiency of Slacking off: Evidence from the Emergency Department." *Econometrica* 86 (3): 997–1030. https://doi.org/10.3982/ECTA13565.

———. 2021. "Influence and Information in Team Decisions: Evidence from Medical Residency." *American Economic Journal: Economic Policy* 13 (1): 106–37. https://doi.org/10.1257/pol.20180501.

Chan, Jr, David C., and Yiqun Chen. 2022. "The Productivity of Professions: Evidence from the Emergency Department." Working Paper 30608. National Bureau of Economic Research. https://doi.org/10.3386/w30608.

Chandra, Amitabh, and Douglas O. Staiger. 2010. "Identifying Provider Prejudice in Healthcare." Working Paper 16382. National Bureau of Economic Research. https://doi.org/10.3386/w16382.

Chen, Alice, and Darius N. Lakdawalla. 2019. "Healing the Poor: The Influence of Patient Socioeconomic Status on Physician Supply Responses." *Journal of Health Economics* 64 (March):43–54. https://doi.org/10.1016/j.jhealeco.2019.02.001.

Chen, Yiqun. 2021. "Team-Specific Human Capital and Team Performance: Evidence from Doctors." *American Economic Review* 111 (12): 3923–62. https://doi.org/10.1257/aer.20201238.

Chernew, Michael, Zack Cooper, Eugene Larsen Hallock, and Fiona Scott Morton. 2021. "Physician Agency, Consumerism, and the Consumption of Lower-Limb MRI Scans." *Journal of Health Economics* 76 (March):102427. https://doi.org/10.1016/j.jhealeco.2021.102427.

Chorniy, Anna, Janet Currie, and Lyudmyla Sonchak. 2018. "Exploding Asthma and ADHD Caseloads: The Role of Medicaid Managed Care." *Journal of Health Economics* 60 (July):1–15. https://doi.org/10.1016/j.jhealeco.2018.04.002.

Chu, Bryan, Ben Handel, Jonathan Kolstad, Jonas Knecht, Ulrike Malmendier, and Filip Matejka. 2024. "Cognitive Capacity, Fatigue and Decision Making: Evidence from the Practice of Medicine." UC Berkeley.

Clemens, Jeffrey, and Joshua D. Gottlieb. 2014. "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?" *American Economic Review* 104 (4): 1320–49. https://doi.org/10.1257/aer.104.4.1320.

Clemens, Jeffrey, Pierre-Thomas Léger, Yashna Nandan, and Robert Town. 2024. "Physician Practice Preferences and Healthcare Expenditures: Evidence from Commercial Payers." Working Paper. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w33090.

Costa-Ramón, Ana María, Ana Rodríguez-González, Miquel Serra-Burriel, and Carlos Campillo-Artero. 2018. "It's about Time: Cesarean Sections and Neonatal Health." *Journal of Health Economics* 59 (May):46–59. https://doi.org/10.1016/j.jhealeco.2018.03.004.

Coudin, Elise, Anne Pla, and Anne-Laure Samson. 2015. "GP Responses to Price Regulation: Evidence from a French Nationwide Reform." *Health Economics* 24 (9): 1118–30. https://doi.org/10.1002/hec.3216.

Coussens, Stephen. 2022. "Behaving Discretely: Heuristic Thinking in the Emergency Department." Columbia University. https://www.dropbox.com/s/6sdlwnniq2unlnk/behaving_discretely_web.pdf?dl=0&unfurl=1.

Cuddy, Emily, and Janet Currie. 2020. "Treatment of Mental Illness in American Adolescents Varies Widely within and across Areas." *Proceedings of the National Academy of Sciences* 117 (39): 24039–46. https://doi.org/10.1073/pnas.2007484117.

———. 2024. "Rules vs. Discretion: Treatment of Mental Illness in U.S. Adolescents." *Journal of Political Economy*. https://doi.org/10.3386/w27890.

Currie, Janet M, and W Bentley MacLeod. 2020. "Understanding Doctor Decision Making: The Case of Depression Treatment." *Econometrica: Journal of the Econometric Society* 88 (3): 847–78. https://doi.org/10.3982/ECTA16591.

Currie, Janet, and Kate Musen. 2025. "Effects of Medicaid Prior Authorization on Child Antipsychotic Prescribing: Spillovers, Information, and Hassle Costs." ASSA 2025 Annual Meeting. San Fransisco.

Currie, Janet, and Jonathan Zhang. 2023. "Doing More with Less: Predicting Primary Care Provider Effectiveness." *The Review of Economics and Statistics*, February, 1–45. https://doi.org/10.1162/rest_a_01290.

Dahlstrand, Amanda. 2024. "Defying Distance? The Provision of Services in the Digital Age."

DeCicca, Philip, Maripier Isabelle, and Natalie Malak. 2024. "How Do Physicians Respond to New Medical Research?" *Health Economics* 33 (10): 2206–28. https://doi.org/10.1002/hec.4879.

Dickstein. 2017. "Physician vs. Patient Incentives in Prescription Drug Choice." http://www.michaeljdickstein.com/posts/2015/9/24/patient-vs-physician-incentives-in-prescription-drug-choice.

Ding, Yu, and Chenyuan Liu. 2021. "Alternative Payment Models and Physician Treatment Decisions: Evidence from Lower Back Pain." *Journal of Health Economics* 80 (December). https://doi.org/10.1016/j.jhealeco.2021.102548.

Doctor, Jason N., Andy Nguyen, Roneet Lev, Jonathan Lucas, Tara Knight, Henu Zhao, and Michael Menchine. 2018. "Opioid Prescribing Decreases after Learning of a Patient's Fatal Overdose." *Science* 361 (6402): 588–90. https://doi.org/10.1126/science.aat4595.

Doyle, Joseph J., Steven M. Ewer, and Todd H. Wagner. 2010. "Returns to Physician Human Capital: Evidence from Patients Randomized to Physician Teams." *Journal of Health Economics* 29 (6): 866–82. https://doi.org/10.1016/j.jhealeco.2010.08.004.

Doyle, Jr., Joseph J. 2020. "Physician Characteristics and Patient Survival: Evidence from Physician Availability." Working Paper 27458. National Bureau of Economic Research. https://doi.org/10.3386/w27458.

Dubois, Pierre, and Tuba Tunçel. 2021. "Identifying the Effects of Scientific Information and Recommendations on Physicians' Prescribing Behavior." *Journal of Health Economics* 78 (July):102461. https://doi.org/10.1016/j.jhealeco.2021.102461.

Eli, Shari, Trevon D. Logan, and Boriana Miloucheva. 2019. "Physician Bias and Racial Disparities in Health: Evidence from Veterans' Pensions." Working Paper 25846. National Bureau of Economic Research. https://doi.org/10.3386/w25846.

Ellyson, Alice M, Jevay Grooms, and Alberto Ortega. 2022. "Flipping the Script: The Effects of Opioid Prescription Monitoring on Specialty-Specific Provider Behavior." *Health Economics* 31 (2): 297–341. https://doi.org/10.1002/hec.4446.

Epstein, Andrew J., Sean Nicholson, and David A. Asch. 2016. "The Production of and Market for New Physicians' Skill." *American Journal of Health Economics* 2 (1): 41–65. https://doi.org/10.1162/AJHE_a_00033.

Facchini, Gabriel. 2022. "Forgetting-by-Not-Doing: The Case of Surgeons and Cesarean Sections." *Health Economics* 31 (3): 481–95. https://doi.org/10.1002/hec.4460.

Fortin, Bernard, Nicolas Jacquemet, and Bruce Shearer. 2021. "Labour Supply, Service Intensity, and Contracts: Theory and Evidence on Physicians." *Journal of Applied Econometrics* 36 (6): 686–702. https://doi.org/10.1002/jae.2840.

Frakes, Michael. 2013. "The Impact of Medical Liability Standards on Regional Variations in Physician Behavior: Evidence from the Adoption of National-Standard Rules." *American Economic Review* 103 (1): 257–76. https://doi.org/10.1257/aer.103.1.257.

Frakes, Michael D., and Jonathan Gruber. 2022. "Racial Concordance and the Quality of Medical Care: Evidence from the Military." Working Paper 30767. National Bureau of Economic Research. https://doi.org/10.3386/w30767.

Freedman, Seth, Ezra Golberstein, Tsan-Yao Huang, David J. Satin, and Laura Barrie Smith. 2021. "Docs with Their Eyes on the Clock? The Effect of Time Pressures on Primary Care Productivity." *Journal of Health Economics* 77 (May):102442. https://doi.org/10.1016/j.jhealeco.2021.102442.

Geiger, Caroline K., Mark A. Clapp, and Jessica L. Cohen. 2021. "Association of Prenatal Care Services, Maternal Morbidity, and Perinatal Mortality With the Advanced Maternal Age Cutoff of 35 Years." *JAMA Health Forum* 2 (12): e214044. https://doi.org/10.1001/jamahealthforum.2021.4044.

Goetz, Daniel. 2023. "Telemedicine Competition, Pricing, and Technology Adoption: Evidence from Talk Therapists." *International Journal of Industrial Organization* 89 (July):102956. https://doi.org/10.1016/j.ijindorg.2023.102956.

Gowrisankaran, Gautam, Keith Joiner, and Pierre Thomas Léger. 2022. "Physician Practice Style and Healthcare Costs: Evidence from Emergency Departments." *Management Science*, October. https://doi.org/10.1287/mnsc.2022.4544.

Goyal, Monika K., Nathan Kuppermann, Sean D. Cleary, Stephen J. Teach, and James M. Chamberlain. 2015. "Racial Disparities in Pain Management of Children with Appendicitis in Emergency Departments." *JAMA Pediatrics* 169 (11): 996–1002. https://doi.org/10.1001/jamapediatrics.2015.1915.

Greenwood, Brad N., Seth Carnahan, and Laura Huang. 2018. "Patient–Physician Gender Concordance and Increased Mortality among Female Heart Attack Patients." *Proceedings of the National Academy of Sciences* 115 (34): 8569–74. https://doi.org/10.1073/pnas.1800097115.

Greenwood, Brad N., Rachel R. Hardeman, Laura Huang, and Aaron Sojourner. 2020. "Physician–Patient Racial Concordance and Disparities in Birthing Mortality for Newborns." *Proceedings of the National Academy of Sciences* 117 (35): 21194–200. https://doi.org/10.1073/pnas.1913405117.

Gruber, Jonathan, Thomas P. Hoe, and George Stoye. 2021. "Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers." *The Review of Economics and Statistics*, May, 1–45. https://doi.org/10.1162/rest_a_01044.

Gupta, Atul. 2021. "Impacts of Performance Pay for Hospitals: The Readmissions Reduction Program." *American Economic Review* 111 (4): 1241–83. https://doi.org/10.1257/aer.20171825.

Hill, Andrew J., Daniel B. Jones, and Lindsey Woodworth. 2023. "Physician-Patient Race-Match Reduces Patient Mortality." *Journal of Health Economics* 92 (December). https://doi.org/10.1016/j.jhealeco.2023.102821.
Hoffman, Kelly M., Sophie Trawalter, Jordan R. Axt, and M. Norman Oliver. 2016. "Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs about Biological Differences between Blacks and Whites." *Proceedings of the National Academy of Sciences* 113 (16): 4296–4301. https://doi.org/10.1073/pnas.1516047113.
Horwitz, Jill, Austin Nichols, Carrie H. Colla, and David M. Cutler. 2024. "Technology Regulation Reconsidered: The Effects of Certificate of Need Policies on the Quantity and Quality of Diagnostic Imaging." Working Paper. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w32143.
Howard, David H., Guy David, and Jason Hockenberry. 2017. "Selective Hearing: Physician-Ownership and Physicians' Response to New Evidence." *Journal of Economics & Management Strategy* 26 (1): 152–68. https://doi.org/10.1111/jems.12178.
Howard, David H., and Jason Hockenberry. 2019. "Physician Age and the Abandonment of Episiotomy." *Health Services Research* 54 (3): 650–57. https://doi.org/10.1111/1475-6773.13132.
Howard, David H., and Ian McCarthy. 2021. "Deterrence Effects of Antifraud and Abuse Enforcement in Health Care." *Journal of Health Economics* 75 (January):102405. https://doi.org/10.1016/j.jhealeco.2020.102405.
Johnson, Erin M., and M. Marit Rehavi. 2016. "Physicians Treating Physicians: Information and Incentives in Childbirth." *American Economic Journal: Economic Policy* 8 (1): 115–41. https://doi.org/10.1257/pol.20140160.
Johnson, Erin, M. Marit Rehavi, Jr Chan David C., and Daniela Carusi. 2016. "A Doctor Will See You Now: Physician-Patient Relationships and Clinical Decisions." Working Paper 22666. National Bureau of Economic Research. https://doi.org/10.3386/w22666.
Kolstad, Jonathan T. 2013. "Information and Quality When Motivation Is Intrinsic: Evidence from Surgeon Report Cards." *American Economic Review* 103 (7): 2875–2910. https://doi.org/10.1257/aer.103.7.2875.
Kowalski, Amanda E. 2023. "Behaviour within a Clinical Trial and Implications for Mammography Guidelines." *The Review of Economic Studies* 90 (1): 432–62. https://doi.org/10.1093/restud/rdac022.
Linder, Jeffrey A., Jason N. Doctor, Mark W. Friedberg, Harry Reyes Nieva, Caroline Birks, Daniella Meeker, and Craig R. Fox. 2014. "Time of Day and the Decision to Prescribe Antibiotics." *JAMA Internal Medicine* 174 (12): 2029–31. https://doi.org/10.1001/jamainternmed.2014.5225.
Ly, Dan P. 2021. "The Influence of the Availability Heuristic on Physicians in the Emergency Department." *Annals of Emergency Medicine* 78 (5): 650–57. https://doi.org/10.1016/j.annemergmed.2021.06.012.
Ly, Dan P., Paul G. Shekelle, and Zirui Song. 2023. "Evidence for Anchoring Bias During Physician Decision-Making." *JAMA Internal Medicine* 183 (8): 818–23. https://doi.org/10.1001/jamainternmed.2023.2366.
McCullough, Jeffrey S., Michelle Casey, Ira Moscovice, and Shailendra Prasad. 2010. "The Effect Of Health Information Technology On Quality In U.S. Hospitals." *Health Affairs* 29 (4): 647–54. https://doi.org/10.1377/hlthaff.2010.0155.
McDevitt, Ryan C., and James W. Roberts. 2014. "Market Structure and Gender Disparity in Health Care: Preferences, Competition, and Quality of Care." *The RAND Journal of Economics* 45 (1): 116–39. https://doi.org/10.1111/1756-2171.12044.
McKibbin, Rebecca. 2023. "The Effect of RCTs on Drug Demand: Evidence from off-Label Cancer Drugs." *Journal of Health Economics* 90 (July):102779. https://doi.org/10.1016/j.jhealeco.2023.102779.
Miller, Amalia R., and Catherine E. Tucker. 2011. "Can Health Care Information Technology Save Babies?" *Journal of Political Economy* 119 (2): 289–324. https://doi.org/10.1086/660083.
Molitor, David. 2018. "The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration." *American Economic Journal: Economic Policy* 10 (1): 326–56. https://doi.org/10.1257/pol.20160319.
Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *The Quarterly Journal of Economics* 137 (2): 679–727. https://doi.org/10.1093/qje/qjab046.

Neprash, Hannah T., John F. Mulcahy, Dori A. Cross, Joseph E. Gaugler, Ezra Golberstein, and Ishani Ganguli. 2023. "Association of Primary Care Visit Length with Potentially Inappropriate Prescribing." *JAMA Health Forum* 4 (3): e230052. https://doi.org/10.1001/jamahealthforum.2023.0052.
Neumark, David, and Bogdan Savych. 2023. "Effects of Opioid-Related Policies on Opioid Utilization, Nature of Medical Care, and Duration of Disability." *American Journal of Health Economics* 9 (3): 331–73. https://doi.org/10.1086/722981.
Newham, Melissa, and Marica Valente. 2024. "The Cost of Influence: How Gifts to Physicians Shape Prescriptions and Drug Costs." *Journal of Health Economics* 95 (May):102887. https://doi.org/10.1016/j.jhealeco.2024.102887.
Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53. https://doi.org/10.1126/science.aax2342.
Olenski, Andrew R., André Zimerman, Stephen Coussens, and Anupam B. Jena. 2020. "Behavioral Heuristics in Coronary-Artery Bypass Graft Surgery." *New England Journal of Medicine* 382 (8): 778–79. https://doi.org/10.1056/NEJMc1911289.
Olson, Mary K., and Nina Yin. 2021. "New Clinical Information and Physician Prescribing: How Do Pediatric Labeling Changes Affect Prescribing to Children?" *Health Economics* 30 (1): 144–64. https://doi.org/10.1002/hec.4182.
Persson, Emil, Kinga Barrafrem, Andreas Meunier, and Gustav Tinghög. 2019. "The effect of decision fatigue on surgeons' clinical decision making." *Health Economics* 28 (10): 1194–1203. https://doi.org/10.1002/hec.3933.
Persson, Petra, Xinyao Qiu, and Maya Rossin-Slater. 2021. "Family Spillover Effects of Marginal Diagnoses: The Case of ADHD." Working Paper 28334. National Bureau of Economic Research. https://doi.org/10.3386/w28334.
Sabin, Janice A., and Anthony G. Greenwald. 2012. "The Influence of Implicit Bias on Treatment Recommendations for 4 Common Pediatric Conditions: Pain, Urinary Tract Infection, Attention Deficit Hyperactivity Disorder, and Asthma." *American Journal of Public Health* 102 (5): 988–95. https://doi.org/10.2105/AJPH.2011.300621.
Sacarny, Adam, Michael L. Barnett, Jackson Le, Frank Tetkoski, David Yokum, and Shantanu Agrawal. 2018. "Effect of Peer Comparison Letters for High-Volume Primary Care Prescribers of Quetiapine in Older and Disabled Adults: A Randomized Clinical Trial." *JAMA Psychiatry* 75 (10): 1003–11. https://doi.org/10.1001/jamapsychiatry.2018.1867.
Sacarny, Adam, David Yokum, Amy Finkelstein, and Shantanu Agrawal. 2016. "Medicare Letters To Curb Overprescribing Of Controlled Substances Had No Detectable Effect On Providers." *Health Affairs* 35 (3): 471–79. https://doi.org/10.1377/hlthaff.2015.1025.
Sacks, Daniel W., Alex Hollingsworth, Thuy Nguyen, and Kosali Simon. 2021. "Can Policy Affect Initiation of Addictive Substance Use? Evidence from Opioid Prescribing." *Journal of Health Economics* 76 (March):102397. https://doi.org/10.1016/j.jhealeco.2020.102397.
Schnell, Molly, and Janet Currie. 2018. "Addressing the Opioid Epidemic: Is There a Role for Physician Education?" *American Journal of Health Economics* 4 (3): 383–410. https://doi.org/10.1162/ajhe_a_00113.
Shapiro, Bradley T. 2018. "Informational Shocks, Off-Label Prescribing, and the Effects of Physician Detailing." *Management Science* 64 (12): 5925–45. https://doi.org/10.1287/mnsc.2017.2899.
Shurtz, Ity, Alon Eizenberg, Adi Alkalay, and Amnon Lahad. 2022. "Physician Workload and Treatment Choice: The Case of Primary Care." *The RAND Journal of Economics* 53 (4): 763–91. https://doi.org/10.1111/1756-2171.12425.
Silver, David. 2021. "Haste or Waste? Peer Pressure and Productivity in the Emergency Department." *The Review of Economic Studies* 88 (3): 1385–1417. https://doi.org/10.1093/restud/rdaa054.
Simeonova, Emilia, Niels Skipper, and Peter Rønø Thingholm. 2024. "Physician Health Management Skills and Patient Outcomes." *Journal of Human Resources* 59 (3): 777–809. https://doi.org/10.3368/jhr.0420-10833R1.

Singh, Manasvini. 2021. "Heuristics in the Delivery Room." *Science* 374 (6565): 324–29. https://doi.org/10.1126/science.abc9818.
Singh, Manasvini, and Atheendar Venkataramani. 2022. "Rationing by Race." Working Paper 30380. National Bureau of Economic Research. https://doi.org/10.3386/w30380.
Tai-Seale, Ming, and Thomas McGuire. 2012. "Time Is up: Increasing Shadow Price of Time in Primary-Care Office Visits." *Health Economics* 21 (4): 457–76. https://doi.org/10.1002/hec.1726.
Van Parys, Jessica. 2016. "Variation in Physician Practice Styles within and across Emergency Departments." *PLOS ONE* 11 (8). https://doi.org/10.1371/journal.pone.0159882.
Van Parys, Jessica, and Zach Y. Brown. 2023. "Broadband Internet Access and Health Outcomes: Patient and Provider Responses in Medicare." Working Paper. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w31579.
Wallis, Christopher J. D., Angela Jerath, Natalie Coburn, Zachary Klaassen, Amy N. Luckenbaugh, Diana E. Magee, Amanda E. Hird, et al. 2022. "Association of Surgeon-Patient Sex Concordance with Postoperative Outcomes." *JAMA Surgery* 157 (2): 146–56. https://doi.org/10.1001/jamasurg.2021.6339.
Wilding, Anna, Luke Munford, Bruce Guthrie, Evangelos Kontopantelis, and Matt Sutton. 2022. "Family Doctor Responses to Changes in Target Stringency under Financial Incentives." *Journal of Health Economics* 85 (September). https://doi.org/10.1016/j.jhealeco.2022.102651.
Wu, Bingxiao, and Guy David. 2022. "Information, Relative Skill, and Technology Abandonment." *Journal of Health Economics* 83 (May):102596. https://doi.org/10.1016/j.jhealeco.2022.102596.
Zeltzer, Dan, Liran Einav, Joseph Rashba, and Ran D Balicer. 2024. "The Impact of Increased Access to Telemedicine." *Journal of the European Economic Association* 22 (2): 712–50. https://doi.org/10.1093/jeea/jvad035.
Zeltzer, Dan, Liran Einav, Joseph Rashba, Yehezkel Waisman, Motti Haimi, and Ran D. Balicer. 2023. "Adoption and Utilization of Device-Assisted Telemedicine." *Journal of Health Economics* 90 (July):102780. https://doi.org/10.1016/j.jhealeco.2023.102780.