NBER WORKING PAPER SERIES

DIAGNOSING PHYSICIAN ERROR: A MACHINE LEARNING APPROACH TO LOW-VALUE HEALTH CARE

Sendhil Mullainathan
Ziad Obermeyer

Working Paper 26168
http://www.nber.org/papers/w26168

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2019, Revised August 2021

We acknowledge grants from the NIH (DP5OD012161, P01AG005842) and the Per-shing Square Fund for Research on the Foundations of Human Behavior. We thank Amitabh Chandra, Ben Handel, Larry Katz, Danny Kahneman, Jon Kolstad, Andrei Shleifer, Richard Thaler, and five anonymous referees for thoughtful comments. We are deeply grateful to Cassidy Shubatt, as well as Adam Baybutt, Shreyas Lakhtakia, Katie Lin, and Advik Shreekumar, for research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care
Sendhil Mullainathan and Ziad Obermeyer
NBER Working Paper No. 26168
August 2019, Revised August 2021
JEL No. C55,D8,D84,D9,I1,I13

ABSTRACT

We use machine learning to study how physicians make decisions, in particular whether to test a patient for heart attack. We build an algorithmic model of patient risk, and identify cases where physician choices deviate from its predictions. To judge who was right, the algorithm or the physician, we use actual outcome data. Three findings emerge. First, many patients are tested who should not be: at typical cost-effectiveness thresholds, 62% of all tests would be cut. Second, many patients who should be tested are not. These patients go on to suffer adverse health events (including death) at rates exceeding clinical guidelines. A natural experiment using shift-to-shift testing variation confirms these findings: increased testing improves health and reduces mortality, but only for high-risk patients. We estimate the under-tested set to be 61% of the size of the tested population. Machine learning, by measuring value at the patient level rather than population level, is crucial for uncovering under- and over-testing. The simultaneous existence of both cannot easily be explained by physician incentives alone, and instead suggests physician errors. Third, we provide suggestive evidence on why physicians err: (i) they use too simple a model of risk, suggesting bounded rationality; (ii) they over-weight salient information; and (iii) they over-weight representative symptoms—those that fit the stereotype of heart attack. Together, these results suggest the need for health care models and policies to incorporate not just physician incentives, but also physician mistakes.

Sendhil Mullainathan
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
Sendhil.Mullainathan@chicagobooth.edu

Ziad Obermeyer
School of Public Health
University of California, Berkeley
2121 Berkeley Way
Berkeley, CA 94704
and NBER
zobermeyer@berkeley.edu

A data appendix is available at http://www.nber.org/data-appendix/w26168

1 Introduction

A patient arrives in the emergency room complaining of chest pain and nausea. Should she be tested for a heart attack (technically, a new blockage in the coronary arteries)? A missed heart attack can have catastrophic consequences, but the test of it is costly and invasive. The choice is not easy, particularly since many benign conditions (like acid reflux) share symptoms with heart attack. To make the choice, the physician must integrate a diverse set of data to predict the risk a patient is having a heart attack. We use machine learning to study these choices and the predictions underneath them. Though we focus on heart attack, our approach applies more broadly, as all testing decisions can be similarly cast as prediction problems (Kleinberg et al., 2015; Kleinberg et al., 2018; Agrawal, Gans, and Goldfarb, 2019)).

Our sample spans all emergency visits over 2010–2015 at a large, top-ranked hospital.[1] For each of these 246,265 visits, we track tests given, resulting treatments, and subsequent health outcomes, encompassing most (though not all) of the data available to physicians. On a random $\frac{2}{3}$ sample of these data, we train an ensemble machine learning model to predict the outcome of testing, using only information available at the time of the testing decision. We do not naively benchmark physician choices against these algorithmic predictions, assuming that they are accurate. Instead, we use the algorithm only to identify (in the remaining $\frac{1}{3}$ hold-out sample) patient subgroups with *potential* inefficiency, where physicians might have made errors. We then look at actual outcomes for these subgroups to test whether mistakes were made, or whether physicians correctly relied on data unavailable to the algorithm.

This approach reveals two allocative inefficiencies in whom physicians choose to test. First, many patients who *predictably* will not benefit from testing are never-

[1] We also repeat much of our analysis in a large sample of nationally representative Medicare claims.

theless tested. We quantify the value of a test here using the treatment benefits it produces (allowing for the fact that the test itself is imperfect), expressed in cost per life-year saved. By this measure, 62% of tests cost more than $150,000 per life year. Algorithmic predictions are crucial in uncovering these wasted tests. Had we instead followed the usual approach of using overall average yields to assess efficiency, we would have concluded that testing as whole was cost-effective, at $89,714 per life year (Weinstein et al., 1996; Sanders et al., 2016). Machine learning is useful for capturing large patient-level heterogeneity.

Second, at the same time, many patients who *predictably* would benefit from testing nevertheless go untested. One sign of this problem is like Abaluck et al. (2016)'s earlier finding that physician choices deviate from a structural risk model: here, physicians do not test many of the patients with high predicted risk. By themselves, though, such deviations do not establish an error as we do not know what the test results would have been for these patients. Physicians may have valid reasons for not testing them; and these may be unavailable in our data (and to the algorithm): how the patient looks, what they say, the results of x-rays or electrocardiograms (ECGs). The problem cannot be solved by imputing outcomes to the untested.[2]

Health outcomes in the untested, though, provide a way to empirically assess these choices. In the thirty days after their visit, high-risk untested (and untreated) patients exhibit the well-known signs of missed heart attack: "major adverse cardiac events" at rates well above existing clinical guideline thresholds for heart attack.[3]

[2]We illustrate using ECGs, typically missing from many medical datasets and effectively an unobserved variable to our algorithm: we only have them for a subset of patients and so exclude them. For that subset, incorporating these waveforms using deep learning decreases predicted risk for 97.5% of patients, and 100% of the highest-risk untested, suggesting predictions are highly confounded for the untested. Despite growing attention to the 'selective labels' problem, similar biases pervade many machine learning applications (Kleinberg et al., 2018; Kallus and Zhou, 2018; Rambachan, 2021).

[3]Such decision rules (e.g., TIMI, GRACE, HEART) are commonly implemented in emergency medicine. We do not take a stance on whether they are physiologically optimal, only that they represent current physician understanding of who should be tested. If physicians use private information

A third of these events lead to death. These patients appear to have indeed been high risk. Still, it is possible physicians recognize this risk but choose not to test because they deem patients unsuitable for invasive treatments. We find evidence to the contrary. For example, a large fraction do not even receive an ECG, a very low-cost, noninvasive test given to any patient with even a small suspicion of heart issues. Physicians simply seem to overlook the risk for these patients.

For more direct evidence of under-testing, we rely on a natural experiment: a patient's arrival time determines staff seen, and staff vary in their tendency to test for heart attack. Conditioning on the visit's hour and day, this provides plausibly exogenous shift-to-shift variation in testing rates.[4] We find that higher-testing shifts do not, on average, produce better health outcomes, indicating so-called 'flat of the curve' health care: more testing yields little return (Fisher et al., 2003). But as before, averages obscure important heterogeneity. Predicted high-risk patients benefit significantly from higher testing: in the subsequent year, heart attacks and arrhythmias are reduced by 41.9% and deaths by 34.7%, making these additional tests are highly cost-effective ($46,017).[5] Under-testing is also quantitatively important. In our preferred set of policy counterfactuals, recall that the over-tested are 62% of all tests; the under-tested would constitute 61% of all tests.

Why do physicians both over- and under-test? Comparing physician decisions to algorithmic predictions suggests several sources of error. We find evidence of bounded rationality—limits in cognitive resources such as attention, memory or computation (Simon, 1955; Gabaix, 2014; Sims, 2003; Gabaix, 2019; Mullainathan, 2002; Bordalo,

in deciding not to test apparently high-risk patients, adverse event rates should be low.

[4]Patients' observable characteristics appear largely balanced across shifts. In addition, realized yield does not meaningfully relate to shift test rates, suggesting unobservables may also be balanced.

[5]These direct results on health rule out an additional concern: our very definition of risk has so far rested on the assumption that treatments following positive tests are useful. But if physicians overtreat, some of those treatments may fail to improve health, inflating our perceptions of under-testing.

Gennaioli, and Shleifer, 2020). The risk model that best predicts physician testing is much simpler than the one which best predicts actual test outcome. By way of analogy, it is as if the physician 'over-regularizes' (Camerer, 2019). We also find evidence that physicians over-weight salient risks (Tversky and Kahneman, 1974; Bordalo, Gennaioli, and Shleifer, 2012), such as those due to demographics and symptoms. Finally, they over-weight symptoms that are representative (sterotypical) of heart attack (Kahneman and Tversky, 1972; Bordalo et al., 2016). For example, patients with chest pain, a salient and representative symptom, are particularly over-tested.

Health care models have long emphasized moral hazard: paying for tests, rather than outcomes, results in too much testing (Arrow, 1963; Pauly, 1968). Recent work has expanded this perspective to include skill differences, comparative advantage, and error (Abaluck et al., 2016; Chan, Gentzkow, and Yu, 2019; Chandra and Staiger, 2020).[6] We extend this literature by providing evidence of large-scale under-testing (not easily explained by incentives), methodologically showing an important role for machine learning, and finally uncovering some potential sources of error.

Moreover, we show that a core prescription of moral hazard models—incentivize high-testers to act like low-testers—can have perverse effects. Low-testers do test fewer low-risk patients (less over-testing), but at the same time they also test fewer high-risk patients (more under-testing). When physicians make systematic prediction errors, incentives to address one inefficiency exaggerate the other. Models and policies must account for such systematic mistakes, analogous to behavioral hazard models of patient errors (Baicker, Mullainathan, and Schwartzstein, 2015).

[6]Abaluck et al. (2016) highlight how errors may produce both under- and over-testing. Chan, Gentzkow, and Yu (2019) show how differences in skill alone, absent incentives, can produce what appears to be over-testing. Chandra and Staiger (2020) focus on comparative advantage: because some health systems specialize and focus on certain tests and conditions, they may appear to over-treat those. There is also a large clinical literature on error and its behavioral sources (Ægisdóttir et al., 2006; Dawes, Faust, and Meehl, 1989; Elstein, 1999; Redelmeier et al., 2001).

2 Context and Framework

2.1 Medical Context

The coronary arteries provide blood flow to the heart, allowing it to pump blood. By 'heart attack' we mean acute coronary syndrome (ACS): a new blockage in those arteries that reduces blood flow and kills a patch of heart muscle.[7] Its consequences can be immediate (e.g., arrythmia, sudden death) and longer-term (e.g., fatigue, heart failure). Randomized control trials have shown two treatments greatly improve mortality and morbidity if delivered promptly: inserting a flexible metal tube into the blocked artery to restore flow ('stenting'), and for severe cases, bypassing the blockage through open heart surgery.[8] Timely treatment, though, requires timely diagnosis, a challenging task in the emergency department (ED). Even life-threatening blockages have subtle symptoms, e.g., a mild squeezing in the chest, shortness of breath, nausea, or weakness—symptoms that also arise from more benign conditions such as acid reflux, viral infections, and muscle strain. Any suspicion of blockage triggers simple, non-invasive tests (the ECG, laboratory tests like troponin) that help estimate the likelihood of blockage and the urgency of the problem. But no test done in the ED can actually diagnose a blockage.

The definitive test for blockage is cardiac catheterization, an invasive procedure carried out in a separate laboratory, distinct from the ED. A cardiologist inserts an instrument into the coronary arteries, squirts in dye, and visualizes the presence and location of blockages via x-ray. If a blockage is found, a stent is inserted to open it, during the same procedure. An alternative testing pathway adds a step before

[7] We will use 'blockage' from now on to refer to ACS, and distinguish it from a broader category of problems causing damage to the heart (which is often used interchangeably with 'heart attack').

[8] See Amsterdam et al. (2014) for a review. Of note, the emergency treatment we study is distinct from the practice of treating patients with more stable, long-standing coronary artery disease, which does not appear to improve either mortality or morbidity (Al-Lamee et al., 2018).

catheterization: 'stress testing.' This increases patients' heart activity (e.g., by exercising on a treadmill or with a drug). If supply is limited by a blockage, this excess demand will be detected via heart monitoring. The advantage of stress tests is that they are less expensive and non-invasive: if negative, an invasive catheterization may be avoidable. The disadvantage is that, if positive, the patient still needs catheterization to deliver the stent, and precious time has been wasted. The proliferation of both tests has been part of the dramatic reductions in rates of missed blockages in the ED. Before widespread testing, miss rates were substantial: between 2 and 11% of blockages went undiagnosed in the ED (see for example (Pope et al., 2000)).

Both tests, though, are costly: thousands of dollars for stress tests and tens of thousands for catheterization, plus overnight observation and monitoring before testing. They also have health risks, particularly catheterization, which is invasive. In addition to a large dose of radiation, it involves injection of dye that can cause kidney failure, a risk of arterial damage, and stroke (Hamon et al., 2008). The decision to test must weigh potential treatment gains against these costs.

2.2 Framework

In our model, patients are drawn from a distribution $f(X, Z)$ where X and Z are feature vectors observed by the physician; only X is recorded in the data. Blockage B occurs with probability $b(X, Z)$, and a test T for blockage yields a positive outcome with $Pr(Y = 1) = p + B(q - p)$, where p and q are the false and true positive rate respectively; we assume $q > p$. We assume patients can only receive a stent S if $Y = 1$. Stenting requires knowing where to place the stent, which physically requires catheterization.[9] Moreover, medical ethics would make treatment without testing du-

[9]For simplicity, we use stenting, the most common method, to denote all treatments. Note that open-heart surgery also requires prior catheterization, to identify suitability and anatomy for surgery.

bious. (More details are in Appendix 2.) Our key Lemma below does not substantively depend on this assumption, but it does greatly simplify exposition. B, T, Y, and S are all binary variables, and testing and stenting cost c_T and c_S, respectively.

Let a patient's health be

$$W = w(X, Z) - B(\eta - S[\tau - \Delta K])$$

so that an untreated blockage reduces health by η. Stenting reduces η by $\tau < \eta$, but its effect depends on K (contraindications), a binary variable capturing known treatment heterogeneity: some patients (e.g., the frail) may benefit less because invasive treatment poses additional health risks. In the model, we assume $K = k(X, Z)$ and that physicians know it. Because patients would not benefit from testing if they do not benefit from stenting, our empirical work will only focus on the $K = 0$ population to identify inefficiencies in testing.[10] To do so, we initially assume that $k(\cdot)$ only depends on X but weaken this assumption in Section 4.3, to allow for $k(\cdot)$ to depend on unobservables. Given that the test is imperfect, it is useful to define $\tilde{\tau} = E[W|S = 1, Y = 1, K = 0] - E[W|S = 0, Y = 1, K = 0]$, the effect of stenting in $K = 0$ patients who test positive, which includes health consequences of stenting and futile treatments delivered to those without true blockage (unlike τ, the treatment effect when $B = 1$ only). Randomized trials of stenting apply to the $Y = 1$ populations and thus estimate $\tilde{\tau}$. We assume $\tilde{\tau} > 0$. Finally, untreated blockage can lead to adverse events after the visit. Denoted by binary A, they occur with probability $\mu + B(\zeta - \phi S)$, so that stenting reduces their occurrence.

Socially optimal testing would maximize expected health net of costs:

$$\max_{S,T} E[w(X, Z)|X, Z - B(\eta - [\tau - \Delta k(X, Z)]S) - c_T T - c_S S|X, Z],$$

[10] We do not assume the $K = 1$ group benefits from stenting. Practically, $K = 1$ may create higher (health) costs of testing, but we omit this for simplicity; it does not change our core results.

where the choice to stent S can depend on X, Z (which includes K) and test result Y.[11] Given this objective, the socially optimal testing rule is

$$\text{Test iff } b(X,Z) > \frac{c_T + pc_S}{q(\tau - \Delta k(X,Z) - c_S)}.$$

In other words, patients should be tested if their risk of blockage is high enough to justify the costs of testing. All patients with $Y = 1$ are then stented. Efficient testing therefore involves testing only the high-risk patients: an (unachievable) ideal would be to test only those with blockage.

Physicians may not make socially optimal choices. First, we assume they derive additional benefit $\nu > 0$ from testing, e.g., they are paid by the test. So they maximize:

$$E[w(X,Z) - B(\eta - \tau S) - c_T T - (c_S + \delta_S k(X,Z))S|X,Z] + \nu T.$$

Second, their judgments may not match actual risk. Specifically, they estimate the probability of a blockage as $h(X,Z)$, so that deviations $h(X,Z) - b(X,Z)$ represent systematic over- and under-estimation of blockage risk. These preferences and beliefs lead physicians to test according to the rule:

$$\text{Test iff } h(X,Z) > \frac{c_T + pc_S - \nu}{q(\tau - \Delta k(X,Z) - c_S)}$$

and again all positive tests lead to treatment ($Y = 1 \implies S = 1$).

To empirically test for such distortions, note that any subset of patients defined by (X,Z) is either above or below the threshold for efficient testing. Those above the threshold are always tested; as they are high-risk, their yield should be sufficiently high. Those below the threshold are never tested; as their risk is low, they should have few adverse events. To establish inefficiencies, therefore, we only need to find

[11] Notice that as we have set it up, testing only affects health through its effects on stenting; it has no direct effect or other indirect value (such as through information generated for later use). We discuss in greater detail how testing affects stenting in Appendix 1.3.

patient pools that are either (i) tested, but have low average yield or (ii) untested, but have high adverse event rates. The following Lemma formalizes this logic.

Lemma 1. *Suppose there exists a set $\mathcal{V}$ of patient characteristics such that*

$$\underbrace{E[T|(X,Z)\in\mathcal{V}]>0}_{\textit{tested}} \textit{ and } \underbrace{E[Y|(X,Z)\in\mathcal{V}]<\frac{c_T}{\tilde{\tau}-c_S}}_{\textit{but low yield}},$$

then $\mathcal{V}$ is called over-tested *and eliminating all testing in $\mathcal{V}$ increases efficiency.*

Suppose instead $\mathcal{V}$ satisfies

$$\underbrace{E[T|(X,Z)\in\mathcal{V},K=0]<1}_{\textit{untested}} \textit{ and } \underbrace{E[A|(X,Z)\in\mathcal{V},K=0]>\mu+\zeta\big(\frac{c_T+pc_S}{q(\tau-c_S)}\big)}_{\textit{but high adverse events}},$$

Then $\mathcal{V}$ is called under-tested *and testing all $K=0$ patients in $\mathcal{V}$ increases efficiency.*

If physician judgments are erroneous, $h(X,Z)\neq b(X,Z)$, then there can simultaneously be both under-tested and over-tested patient subsets. But if accurate $h(X,Z)=b(X,Z)$, there can only be over-tested subsets, and this happens only if $\nu>0$.

Proof. Consider a set of patients $\mathcal{V}$, and define $\bar{Y}=E[Y|(X,Z)\in\mathcal{V},T=1]$ and $\bar{A}=E[A|(X,Z)\in\mathcal{V},T=0,K=0]$. First, suppose $\mathcal{V}$ satisfies the conditions for being over-tested. If we were to stop testing all patients in $\mathcal{V}$ who are tested, the gain is the savings in c_T for all those who would have been tested. The loss, however, is only those who no longer get treated as a result. Since only the $Y=1$ get treated, this means that at most $\bar{Y}$ of these patients get treated and create treatment benefits $\tilde{\tau}-c_S$. Thus the condition that $E[Y|(X,Z)\in\mathcal{V}]<\frac{c_T}{\tilde{\tau}-c_S}$ ensures that gains outweigh losses.

Now suppose that $\mathcal{V}$ satisfies the conditions for being under-tested, and we were to test all $K=0$ untested patients in $\mathcal{V}$. Given the optimal testing rule, for the $K=0$ patients, it is optimal to test these patients if $b(X,Z)>\frac{c_T+pc_S}{\tau-c_S}$. Given that $\bar{A}=\mu+\eta(b(X,Z))$, it is optimal to test these patients if $\bar{A}>\mu+\eta\left(\frac{c_T+pc_S}{\tau-c_S}\right)$, which is the condition for being under-tested.

Finally, if we assume $b(X,Z) = h(X,Z)$ the physician testing rule above becomes

$$\text{Test iff } b(X,Z) > \frac{c_T + pc_S - \nu}{q(\tau - \Delta k(X,Z) - c_S)}$$

and if $\nu > 0$ can only produce over-testing. If $h(X,Z) \neq b(X,Z)$, it is clear that any kind of over- or under-testing is possible since $h(X,Z)$ can be set to any value. □

Four points are worth noting about this Lemma. First, it illustrates the role of machine learning in our analysis: it serves only to identify *candidate* subsets $\mathcal{V}$ where inefficiencies might be present. Second, once identified, inefficiencies are evaluated using available outcome data: there is no imputation of outcomes. Instead, the key calculations rely only on measured quantities: yield Y for the tested and adverse events A for the untested. Similarly, the relevant thresholds can be calculated using the clinical literature, as we describe in detail below.[12] Third, it allows the physicians to have access to information (Z) that the algorithm does not: it holds for subsets $\mathcal{V}$ identified using only X. Finally, the Lemma links the evidence to an underlying model of physician behavior. Moral hazard alone (bad incentives) can only produce over-testing but not under-testing; mis-prediction, however, can produce both.

It is useful to contrast this model with two others. Chan, Gentzkow, and Yu (2019) model radiologists who receive a noisy signal about patient risk and choose a diagnostic threshold.[13] While superficially analogous to $h(X,Z)$ and ν, a crucial difference is that in their model physicians are aware their signal is noisy (and compensate for it by testing more, to reduce their miss rate). Physicians in our model are unaware of their errors and view their predictions as correct. Our model is closest to Abaluck et al. (2016), who also model physician error. The key difference with them is in how

[12]The adverse event threshold in the Lemma cannot be easily stated in terms of model primitives (i.e., the risk of blockage, the imperfect performance of testing, the impact of treatment on the health) because several key parameters (i.e., p, q, μ, ζ) are unknown.

[13]Norris (2019) makes similar points in a model of judicial decision-making.

we characterize under-testing: we do not assume $b(X, Z)$ is accurate, or define under-testing as deviations of decisions from risk. Instead, we assume measured health outcomes reflect undiagnosed blockage and use these to characterize under-testing.

3 Data and Methods

Our primary data come from the electronic health record (EHR) of a large urban hospital from January 2010 to May 2015. It is an academic medical center, consistently ranked in the top 10 best hospitals in the country and affiliated with a top-ranked medical school, thus widely believed to provide high-quality care. We begin with all visits to the ED in that period, then exclude patients 80 years or older, those with poor-prognosis like known metastatic cancer or dementia, those with hospice or nursing home care, those with a known recent blockage (or treatment of one), and those who died in the ED before they could be sent for testing.[14]

We choose not to exclude those with seemingly obvious non-cardiac symptoms to avoid potentially arbitrary judgments. While some cases are clear (e.g., an ankle sprain), many are not: blockage can present in diverse ways. Worse, we do not observe all of a patient's symptoms, only the one judged most important by the triage nurse.[15] Instead, we use the full sample, and include recorded symptoms in our predictor to make it an empirical question. By including cases highly unlikely to be a blockage, the algorithmic prediction task does become harder: very high-risk cases are comingled with (effectively) zero-risk patients. If it fails, it will appear as an inability to separate high-risk patients from less risky ones. Our final sample has 246,265 ED visits (indexed by j), by 129,859 patients (indexed by i).

[14] See Shanmugam et al. (2015) and Obermeyer et al. (2017) for rationale and details.

[15] Appendix Table A.17 shows the presenting symptom for those ultimately found to have blockage. Non-obvious symptoms (e.g., foot and ankle complaints, nose bleed) are rare but present.

3.1 Definitions of Key Variables

In this sample, we define testing $T_{ij} = 1$ if patient i on has procedure codes for either stress testing or catheterization in the 10-day window (inclusive) following visit j.[16] We define treatment $S_{ij} = 1$ if there is a procedure code for stenting or open-heart surgery (CABG) in the 10-day window following the visit.

To define test yield Y_{ij}, we rely on the principle that a positive test always leads to stenting: a cardiologist should not subject a patient to the risks of emergency catheterization unless she has already decided the patient would benefit from a stent (if a blockage is detected). So we set $Y_{ij} = S_{ij}$ for the tested. As we describe further in Appendix 1.3, physicians may over-treat conditional on test results (e.g., because of moral hazard, or false-positive tests). One might worry this by itself could artificially produce our results. It does not for two reasons. First, over-testing is established through low yield. If physicians over-treat, yield will be too high, making it less likely we find over-testing. Second, establishing under-testing does not use information on the yield of testing—only health outcomes—and hence is unaffected.

To flag patients with contraindications $K_{ij} = 1$, we first observe whether they show evidence of poor health prior to visit j (see above). Second, we observe whether they have explicit evidence of damage to heart muscle (which we will use to denote the medical concepts of infarction and ischemia), a broad category of heart problems including blockage, at the end of visit j: physicians can note such diagnoses (which is financially incentivized), or we can observe a positive troponin laboratory test which is suggestive of such problems. If either are present, we assume the physician was aware of possible blockage, but decided not to pursue it further because of a contraindication. This assumes all contraindications are measured in our data. In Sec-

[16] We collapse these two tests into one for simplicity (as is reflected in our model). Treating the two tests separately does change our results materially. In Appendix 4, we show the results of performing counterfactuals for each test separately, e.g., eliminating all stress tests.

tion 4.3, we explore a broader set of contraindications unobserved in our data but observed by the physician.

Cost effectiveness is calculated using parameters and assumptions from the literature, summarized in Mahoney et al. (2002) and described in more detail in Appendix 3. We first define η_{ij} to be the life-years a patient would lose from a blockage, both fatal and non-fatal (the latter using a standard discount rate for quality of life losses), based on the patient's age and basket of pre-visit observed chronic illnesses. Clinical trials provide estimates of *average* gains from timely treatment, $\tilde{\tau}$. The most relevant trials, from which we draw our estimate of a 25% reduction in mortality and morbidity, randomize testing pathways, e.g., immediate vs. delayed catheterization.[17] We conduct a sensitivity analysis using a wide range of plausible estimates in Appendix 3. We then set $\tilde{\tau}_{ij} = \eta_{ij}\tilde{\tau}$, life years saved by treatment, for patients with $K_{ij} = 0$. We will not explicitly define $\tilde{\tau}_{ij}$ for $K_{ij} = 1$. We account separately for the financial costs of testing c_{Tij} and treatment c_{Sij}, by type of testing and treatment received. Together, this provides a cost per life-year saved by testing a given patient set. To define over-testing, we use a cost-effectiveness threshold of $150,000 per life-year.

We form an indicator $A_{ij} = 1$ if a patient i experiences a 'major adverse cardiac event' after visit j within a short time window (30 days). The intuition is that blockages have consequences—indeed, this is why we test and treat blockages—that manifest shortly after onset. We draw on clinical literature that defines these events using the EHR, in a way that shows excellent agreement with expert judgment after chart review (e.g., Wei et al. (2014)). These events fall into three categories: delayed diagnosis and treatment of blockage and diagnosed damage to heart muscle, which we confirm with laboratory biomarkers (positive troponin); malignant arrhythmia, which we

[17] Given our assumption that positive tests are always treated, this is equivalent to the definition of $\tilde{\tau}$ we lay out in the model.

measure using diagnosis codes and cardiopulmonary resuscitation procedures; and mortality, which we obtain via linkage to state Social Security data. Importantly, apart from mortality, adverse events are only measured if the patient returns to the same health system we study for care. So A_{ij} may be a lower bound on true adverse event rates, relative to widely accepted thresholds from studies that perform active follow-up of enrolled patients. To define objective thresholds for levels of risk that would mandate consideration of testing for blockage, we rely on widely implemented decision rules (e.g., the HEART score of Backus et al. (2010)), supported by recommendations from professional societies: 2% over the 30 days after visits. We do not assume such thresholds are optimal; rather, we assume that physicians believe them to be optimal, and thus would not knowingly leave high-risk patients untested. More details are in Appendix 1.3, and we provide additional justification of this threshold based on cost-effectiveness in Appendix 3.2.

Table 1 shows that the overall rate of testing is 3% of all visits (1.3% with immediate catheterization and 2% with stress tests, of which 0.3% subsequently had catheterization, implying a positive stress test). Table 2 shows that, among the tested, the rate of treatment is low: 14.6% (12.9% with stents and 1.8% via open-heart surgery). Among the untested, 27.5% and 11.1% have an ECG and troponin performed, respectively, indicating suspicion for blockage; 1.2% and 1.9% have explicit evidence of damage to the heart, via the physician's diagnosis ex post and a positive troponin test, respectively. 1.1% had 30-day adverse events.

3.2 Algorithm Design

Our machine learning estimator of risk $\widehat{m}(\cdot)$ is an ensemble model that combines gradient boosted trees and LASSO. It takes as its input vector X_{ij}, 16,405 characteristics

of patient i, observable at the start of visit j.[18] This includes patient demographics; diagnoses, procedures, laboratory results, and quantitative vital signs, measured over the two years prior to the visit; and the symptom recorded at the ED triage desk at the start of the visit. We train estimator $\widehat{m}(X_{ij})$ to predict the yield of testing Y_{ij} among the tested, as a close proxy for risk of blockage at the time of an ED visit.[19] To leverage risk information contained in the much larger set of untested patients, we also use predictions on adverse events $A_{ij} = 1$ among untested patients as inputs to the model predicting Y_{ij}. Training happens in a random 75% sample of patients, and all results below are shown in the remaining 25% hold-out set, except where noted. We split our dataset at the patient, not the observation, level, so that all visits from a given patient are assigned to either the training or hold-out set. More details can be found in Appendix 2. While we cannot share patient-level information to protect privacy, our code repository is publicly available on GitLab [link].

We emphasize that Lemma 1 is valid even if the algorithm is inefficient (or even inconsistent) since it applies to all subsets, however identified. Inefficient algorithms may fail to find under- or over-tested subsets if they do exist but if they find one that satisfies the inequalities then it will be an inefficiency, irrespective of the algorithm's accuracy. It should be added that even a "perfect" algorithm where $m(X) = E[Y|X]$ may fail to find all inefficiencies because it does not have access to Z and so may (for example) miss physician errors involving Z.

[18] We carefully form these variables so that they contain only information available to the physician at the time of the decision. Practically, this means we include no information after triage, on arrival to the ED; we do not include physician notes (which can be completed after the visit) or any data collected later in the course of the ED visit.

[19] To streamline terminology, we will refer to this quantity as 'predicted risk.'

4 Results

4.1 Over-testing

The top panel of Figure 1 shows how well our risk model predicts the outcome of testing. In the hold-out set, we sort tested patients into deciles based on predicted risk. For each decile (x-axis), we calculate the realized yield of testing (y-axis).

The numbers from this figure are also shown in Table 3. Comfortingly, realized yield rises with predicted yield. The algorithm also produces a wide dispersion in realized yields—from 0.01 yield in the lowest decile to 0.55 in the highest decile.

The bottom panel of Figure 1 converts these yields into cost-effectiveness. As in the top panel, patients are sorted by predicted risk, but this time into quintiles (x-axis).[20] The y-axis now shows the implied cost-effectiveness of testing patients in a quintile, in units of thousands of dollars per life year. The y-axis shows a commonly used threshold for judging cost-effectiveness, $150,000, as well as the cost-effectiveness of selected other procedures for comparison. This figure illustrates a great deal of inefficient testing. The bottom quintile of all tests are extremely cost-ineffective: $1,352,466 per life year. For comparison, biologics for rare diseases (some of the least cost-effective technologies that health systems sometimes pay for) are typically estimated at around $300,000 per quality adjusted life year.[21] Even the second quintile is very cost-ineffective at $318,603 dollars per life year.

With these data, we can calculate a precise policy counterfactual as described in Lemma 1: dropping individual tests whose cost effectiveness predictably falls below a threshold. For example, at a $150,000 life-year valuation, we would drop 62.4% of

[20] We use larger bins here because the denominator depends on the yield rate, which approaches zero in the lowest risk patients, leading to noisy estimates in smaller bins.

[21] Appendix (3) shows that these estimates are not sensitive to the particular choice of parameters in our analysis, and in particular hold over wide ranges of possible treatment effect sizes.

the lowest-value tests, with a combined cost-effectiveness of $265,114 per life-year. These results only deal with one kind of counterfactual: eliminating the particular tests physicians decided to do (i.e., stress tests or catheterizations) on patients in a given predicted risk bin. Since we have two types of tests, Appendix 4 also explores other counterfactuals. A notable finding is that stress testing (as opposed to catheterization) is so low-value that eliminating it altogether would improve welfare, as has been previously suggested (Prasad, Cheung, and Cifu, 2012). Taken together, the results in Figure 1 and these policy counterfactuals suggest a great deal of over-testing.

4.2 Under-testing

At the same time, testing in the high-risk bins appears very cost effective: in columns (1) and (2) of Table 3, we see the highest-risk quintile of tests cost only $46,017 per life year, comparable to cost-effective interventions like dialysis. In Column (3), we show the testing rates by risk bin for all patients in the hold-out set (for comparability, these bins continue to be formed using the same quintile cutoffs we used for the tested). We see that physicians do test higher-risk patients more. But strikingly, a large fraction of high-risk patients go untested—only 38.3% are actually tested.

Of course, this only tells us that the physician and the model disagree, not who is right.[22] The physician has access to a host of information unavailable to the model: how the patient looks, what they say, or crucial data such as x-rays or electrocardiograms (ECGs). These data elements are likely to be predictive of yield; if they are also predictive of testing, this private information will create selection bias, causing

[22]To some extent, any two models of risk—even very good ones—may differ due to noise. So perhaps any discrepancies we see between the physician and the model could simply be the consequence of comparing two well-fit models to each other. In Appendix Figure A.11, we compare two machine learning models fit on separate samples of our training set, and find these correlate much more strongly than the model and the physician do. More importantly, we perform a variety of tests below, that directly test for error, both in the sense of welfare-enhancing counterfactuals, and specific behavioral errors.

untested patients to have far lower yield than predicted based on observables.

Because we lack test results on the untested, we have no way to quantify the magnitude of the problem. But a simple calculation suggests a large bias. The hold-out set has 266 positive tests; taking model predictions at face value would imply ten times as many positives (2,738) were we to test the predicted high-risk untested, implausibly large. To show the role of private information more directly, Appendix 5 incorporates data from ECGs, observed by the physician but not routinely observable in health datasets, into risk predictions.[23] For patients with ECG data available, we show that several ECG features (e.g., ST-elevation, ST-depression) predict both the physician's test decision and the yield of testing: physicians are using these data effectively. We then directly incorporate the ECG waveform into new risk predictions, via a deep learning model. This decreases model-predicted risk for 97.5% of patients, and 100% of the highest-risk untested. So the model without the ECG was significantly over-estimating the risk of the untested patients.

So following Lemma 1, we look for evidence of untreated blockages in the form of adverse events in the 30 days after visits. Among all eligible untested patients, the rate of adverse events is 1.1%, well below the 2% clinical threshold, implying that (reassuringly) testing all untested patients does not make sense.[24] Figure 2 shows these adverse event rates (y-axis) by deciles of predicted risk. For comparability, the Figure uses bin cutoffs for deciles formed in the tested, so that the bins are of unequal sizes in the untested: in particular, because the untested are lower-risk than the tested, bin sizes decrease in risk. Panel (a) shows all adverse events, (b) shows diagnosed blockage or arrythmia, and (c) shows death. Patients in the highest-risk

[23] Since not all patients have ECGs, even in our data it cannot be used in our main algorithm.

[24] In Appendix Figure A.2 we show that the 2% adverse event threshold used here in the untested aligns (approximately) with the cost-effectiveness thresholds we used in the tested: patients whose predicted risk gave them a cost-effectiveness of $150,000 per life year when tested have an adverse event rate of at least 3.4% when untested.

decile have very high 30-day adverse event rates. For example, the highest-risk bin contains 0.15% of the untested, who go on to have an adverse event rate of 15.6%. The second highest-risk bin contains 0.75% of the untested and has adverse event rate of 6.81%; together the top two bins have an adverse event rate of 8.26%. In fact, the crossover point where the adverse event rate becomes statistically indistinguishable from the 2% threshold is the 6^{th} risk bin, which means that the top 4 bins (which comprise 6.9% of the untested) all have high enough adverse event rates that they would merit consideration for testing under current guidelines.

These adverse events are not simply billing codes, which might exaggerate the incidence of actual health problems, due to incentives to over-test or treat. These codes have been confirmed with biomarker evidence of damage to the heart muscle, in the form of troponin laboratory results. Panel (b) shows the rate of these diagnosed, confirmed events in the highest-risk bin: 4.9%. Panel (c) also shows 30-day mortality, indicating that the highest-risk bin experiences death at a rate of 3.3%—over one-third (45%) of all adverse events in this bin. These data alone suggest a great deal of under-testing. However, there is a potential confound, which we address next.

4.3 Accounting for Differences in Treatment Benefits

These high adverse event rates establish that predicted high-risk patients who go untested are indeed high-risk. But it does not establish that failing to test them was a mistake. Adverse events rule out private information by physicians about risk, but not private information about the suitability of treatment. It is possible that physicians recognized these patients as being high-risk, but also recognized them as having lower return to treatment, and chose not to test them for that reason. In particular, we may have mismeasured K_{ij}. In excluding patients $K_{ij} = 0$ from our sample (by excluding those with prior ill health, and by excluding untested patients

in whom the physician appears to suspect heart problems), our measure K_{ij} may have failed to capture other elements of K that the physician observes. One fact suggests that these unobservables are not large: the average age of the untested we flag for testing is 58.5 (fairly close to the mean age of the tested, 57.8), while the average age of those with observed contraindications is 68.5. At least on this crucial observable, the high-risk untested are more like the tested than the too frail to test.

To address this problem more thoroughly, we use a clinical fact. When physicians suspect a blockage, even if the patient is ineligible for testing or treatment, there are still important actions they can and must take. At a minimum, everyone the physician suspects of a blockage will be given an ECG—a low-cost, non-invasive test. Even for treatment-ineligible patients, the ECG guides medications (e.g., blood-thinners) and decisions about intensity of monitoring (e.g., whether to admit to the ICU). Similarly, the troponin blood test is also given as it provides critical information on the nature and extent of any blockage. So if we remove patients with an ECG or troponin from our calculations, we will have removed all patients the physicians had suspected at all of a heart problems, leaving us with a pool of unsuspected patients.[25] Among the remaining unsuspected pool, we then recalculate the adverse event rate. If the high adverse event rates in the whole population are due to physicians *knowingly* leaving some high-risk patients untested, because they are unsuitable for treatment, then this unsuspected pool should have a low adverse event rate; and specifically the rates should be below the clinical threshold for testing.

The top two panels of Figure 3 first show the fraction of patients who did not receive an ECG (Panel a) or troponin (Panel b) by quartile of predicted risk. As expected, higher-risk patients are on average perceived as such by physicians: they are

[25]Because some patients are given ECGs and troponins for other reasons, this approach produces a *lower bound* on the extent of under-testing (it removes treatment ineligible patients but also others).

less likely to lack one of these tests. Though decreasing, the fractions remain substantial in the highest quartiles: 27.4% lack an ECG (vs 78.6% in the lowest-risk bin), and 59.2% lack a troponin result (vs 94.4% in the lowest-risk bin). The bottom two panels show the adverse event rates in only these patients without an ECG or without troponin. We find here that for the high-risk patients without such tests, adverse event rates remain high. For patients in the highest-risk bin, the realized adverse event rate is 4.3% in those without an ECG, and 6.6% in those without a troponin. These rates are 3.2 percentage points (SE: 1.3 p.p.) and 1.2 percentage points (SE: 1.1) lower than the 7.5% rate in the full population, respectively, but they still remain significantly above the clinical threshold for testing of 2%.[26] Together, these results suggest that physicians *do* have private information both about the risk of blockage and about suitability for treatment, but that even after accounting for them, there is still substantial under-testing.

4.4 Natural Experiment

While these data provide clear evidence of under-testing, this evidence is indirect, based on clinical thresholds. It would be reassuring to have more direct evidence that testing these neglected high-risk patients would impact their health. Ideally, we would measure the impact of testing some high-risk patients at random, and see if in fact mortality and long-term adverse event rates decrease significantly.[27] While such

[26]Appendix 6.3 describes another sensitivity analysis, in which we eliminate patients who were admitted to the hospital with an uncertain diagnosis (e.g., those with a symptom-based diagnosis code like 'chest pain,' as opposed to a specific disease), in whom physicians may have latent concern for blockage. When we calculate adverse event rates in the remaining patients—those in whom the physician felt sure enough to assign an alternative diagnosis other than blockage, and those discharged home from the ED and thus at very low risk of serious problems—we find similar results: a rate of adverse events 8.43% in the highest-risk bin, as opposed to 8.26% in the full population.

[27]In the context of the framework, the natural experiment measures $E[W|T=1, m(X)] - E[W|T=0, m(X)]$ but only for the marginal patient physicians test; we can then see whether these returns are above or below what would merit testing.

an experiment is beyond the scope of this paper, we can exploit natural variation in our data that might serve as a (limited) proxy for it.

When a patient arrives at the ED, they are seen by a team of providers, largely nurses, at the triage desk. As Chan and Gruber (2020) note, the triage process can influence downstream decision making by physicians regarding testing. For example, a nurse can notice that a patient with chest pain is sweaty, or not; she can ascribe it to the hot and humid weather, or not; and she can share her impressions with the physician when he brings the patient back into the room. As a result, we hypothesized that the testing rate, while ultimately determined by the physician, could be affected by the particular make-up of the team working the triage desk. As who is present varies over time, this creates a 'natural experiment' based on the exact time a patient showed up; and as shifts are not perfectly synchronized with the calendar, we can control for day of week and hour of day.

Our data do not track the exact identity of the triage team, but we do know the times at which shifts begin and end. This lets us calculate the average testing rate of all other patients seen on a shift, $\bar{T}_{-j}$, to instrument for whether patient-visit j is tested. For this to be a valid instrument, we assume that (i) the triage shift affects long-term health outcomes only through testing, and (ii) that patients are balanced on unobservables across shifts; we discuss both assumptions below. We perform this analysis on a slightly different sample than used so far. To maximize power, we use the full dataset, not just the hold-out. To avoid over-fitting, we use 5-fold cross-validation to predict risk. In addition, to address non-independence of health outcomes across visits, we restrict the sample to each patient's first visit. [28]

Overall, there is reasonable variation in likelihood of testing across shifts: for

[28]Results restricted to the hold-out are very similar, just less precise as we would expect given the sample size. We also check that results are similar if we include all visits and cluster standard errors, but prefer this specification for its transparency.

example, a patient in the highest-risk bin arriving on a Monday evening is 18% more likely to be tested by the highest- (19.9%) vs. lowest-decile (16.8%) shifts. Regressing an individual's test (T_{ij}) on the shift's (leave-one-out) testing rate ($\bar{T}_{-j}$), controlling for time fixed effects (year, week of year, day of week, and hour of day) and patient risk, we find that a one standard-deviation increase in shift testing rate (2.3 percentage points) increases individual testing probability by 0.19 percentage points (SE: 0.06), or 6.7% of the base test rate (see Appendix Table A.12).[29]

Figure 4 shows how patient observables are balanced across shifts. The top Panel shows the results of regressing pre-triage variables X_{ij} on the shift testing rate. We do find statistically significant differences in predicted risk across triage testing rates (p=0.051), but they are very small in magnitude: a 1 SD increase in $\bar{T}_{-j}$ implies 0.007 SD difference in predicted risk. We find no statistically significant difference for fraction in each predicted risk bin, nor age, sex, self-reported race, income, or risk factors for heart disease. Together, these results suggest that observables are (largely) balanced across shifts. In the bottom panel, we plot for each shift, the average testing rate for all patients who arrive in that shift (in percentile terms, x-axis) and the average predicted risk of those patients (y-axis). We see that at every level of testing rate, there is large variability in predicted risk.

In Appendix Table A.12, as another test for balance, we regress test T_{ij} on predicted risk and its interaction with $\bar{T}_{-j}$. If patients in high-testing shifts are riskier on unobservables, they should have higher yield than expected based on risk, leading the interaction term to be positive. In fact, there is no significant interaction. While estimates are imprecise, they do argue against large imbalance on unobservables.

[29]In Appendix Table A.11, we also rule out that hospital capacity constraints on testing facilities might be reducing the likelihood of testing, by showing that a visit's likelihood of testing is not affected by the number of tests done in the 24-72 hours before the visit.

We then measure the overall impact of testing on health by estimating:

$$A_{ij} = \beta_0 + \bar{T}_{-j}\beta_1 + \widehat{m}(X_{ij})\beta_2 + \mathsf{TimeControls}_j\beta_3 + \epsilon_{ij}. \quad (1)$$

That is, we regress adverse event rates on shift testing rates, controlling for time fixed effects (year, week of year, day of week, and hour of day) and patient risk. The top panel of Table 4 shows the results. On average, there is no effect of being seen in a high-testing shift. Neither diagnosed adverse events from day 31–365 after visits (Column 1) nor death, whether measured over the same period as diagnosed events (Column 2) or over the full year after visits (Column 3) are affected.[30]

As before, the average effect may conceal a great deal of heterogeneity: we found under-testing only in high-risk patients. So we re-estimate (1), but include an interaction term $\bar{T}_{-j} \times \widehat{m}(X_{ij})$ to Equation 1; this allows the effect of shift to vary by predicted risk. The results of this regression are in Table 4. We find a large and significant negative interaction term, indicating lower rates of diagnosed events and death in higher-risk patients. This confirms that physician private information about treatment heterogeneity cannot account for our findings: increased testing improves health. It also provides some reassurance regarding the exclusion restriction in our experiment: if triage affected long-term outcomes in ways unrelated to testing for blockage, we would expect to see broader effects, not just among the predicted high-risk for blockage. We emphasize that this does not imply that *all* high-risk untested patients would benefit from testing: we are constrained by the extent of variation in testing rates in our quasi-experiment, and can say nothing about patients who are never tested (i.e., even in the highest-testing shifts).

These coefficients can be used to simulate, at the margin, the benefit of testing

[30]We measure some outcomes over the 30–365 days after ED visits because tested patients are mechanically more likely to be diagnosed with heart problems than untested patients, simply by virtue of being in the hospital for testing. By contrast, our mortality data come from linkage to Social Security data, and so do not suffer from this difference in ascertainment.

high-risk patients. To simulate such a counterfactual, we first use random-effects model of testing rate by shift to parametrize the shift-to-shift variability.[31] We then bin shifts into quartiles based on their random effect, allowing us to simulate higher- vs. lower-testing regimes in our data (from a 5.8% chance of testing in lowest quartile to 32.3% in the highest). We then simulate moving the riskiest quintile of patients from their observed shift quartile to the highest-testing shift quartile (those already in the highest bin are left unchanged). We then apply the coefficients in Table 4 to estimate the effect on mortality. This policy would lead to increase tests equal to 0.48% of the untested (or 15.6% of the current rate) and counterfactual one-year mortality would be 2.5 p.p. lower, or 32% relative to the base rate of 7.7%.

As a group, the evidence so far reveals high-risk untested patients who ought to have been tested. First, they have high adverse event rates of the kind that suggest undiagnosed blockage. Second, physicians do not appear to have recognize the blockage: many were not given simple tests given to everyone suspected of any heart trouble (ECG or troponin); and these also had high adverse event rates. Finally, we see in the natural experiment, that plausibly exogenous increases in testing improve health and only where we would expect: in the high-risk. Each finding has its limitations, but together, they make the case that testing high-risk untested patients would increase welfare as strongly as possible without a randomized trial.

4.5 Nationally Representative Data

These results come from a single hospital. To check their generality, we replicate them in a nationally representative 20% sample of Medicare fee-for-service patients, from January 2009 through June 2013. These data are limited in several important

[31] The leave-one-out shift testing rate, while useful for identification of the effect of testing, does not capture the full variation in observed testing rate across shifts. Appendix 7.3 contains more details on the model, which controls for the same vector of time variables and patients' predicted risk as above.

ways. Because they are based on insurance claims, not EHR data, they contain very limited patient information. For example, we do not have ECGs, lab values, or other biological measures, nor do we have arrival time and shift timing data that would let us recreate our natural experiment. These caveats aside, these data do let us replicate our estimates of over- and under-testing from Sections 4.1 and 4.2. Applying similar exclusions to those used in the single-hospital data, we arrive at a final sample of 4,425,247 Medicare visits by 1,602,501 patients, of whom 4.4% were tested. Of the tested, 12.4% received treatments. Of the untested, 5.3% had 30-day adverse events.

In Figure A.7, we see that yield of testing and cost-effectiveness both increase in predicted risk (as in Figure 1), with many tests being predictably cost-ineffective. We also find many high-risk untested patients with adverse event rates exceeding clinical thresholds. Figure A.8 shows that 3.8% of the highest-risk patients are diagnosed with an adverse event, and an additional 1.5% die. In summary, we find both over-testing (52.6% of all tests) and under-testing (at least 17.9% of the tested).[32]

5 Why do Physicians Make Testing Errors?

We have shown that physicians mis-predict, testing predictably low-risk patients and failing to test predictably high-risk patients. In this section, we try to better understand *how* physicians mis-predict. To do so, we examine how physician testing decisions deviate from predicted risk. Our approach builds on a long tradition of research

[32]Lacking a credible quasi-experiment in these data, we instead rely on a conservative lower-bound for under-testing: we assume that the realized adverse events in predictably high-risk untested patients lower-bounds the under-tested population. We consider this conservative because it assumes that under-testing is concentrated in the smallest possible number of patients, all of whom would have ex ante probability 1 of an event. This may be one reason for the smaller level of under-testing here. Another may be the nature of claims data: low-risk tests may be easy to identify with claims, while high-risk misses may require the richer EHR data. An important caveat to all these results is that we do not observe ECG or troponin testing, so we do not have the same ability to identify countraindicated patients on the basis of observables.

comparing clinical judgment to statistical models as a way to gain insights into decision making, often amongst physicians (Ægisdóttir et al., 2006; Dawes, Faust, and Meehl, 1989; Elstein, 1999; Redelmeier et al., 2001), as well as the clinical literature on diagnostic error (Croskerry, 2002; Graber, Franklin, and Gordon, 2005; IOM, 2015). We view this as exploratory: a way to shed light on potential psychology at work, rather than to structurally estimate a specific model of physician decisions.

5.1 Boundedness in Physician Judgments

One reason physicians may make errors is that the optimal risk model is quite complex: our own machine learning model uses 16,405 variables. Bounded rationality may lead them to use a simpler approximation. Such simplification is analogous to regularization in machine learning (Camerer, 2019). To avoid over-fitting, algorithms do not pick the model that fits best in sample. Instead they estimate a best-fit model for each level of complexity. Then a complexity level is chosen by asking which of these best-fit models produces best out-of-sample fit. To study physicians, we use this set of best-fit models for each complexity. But we now ask which model complexity best predicts physician choices, not out-of-sample risk. If physicians are boundedly rational, the model that best predicts physician choices should be simpler than the one that best predicts actual risk.

We implement this procedure using the LASSO model of risk, one component of our full ensemble model, because it has straightforward measure of complexity: the number of non-zero coefficients included in its linear model.[33] For $k \in [0, 1500]$ we thus retain the set of best-fit LASSO models that has exactly k non-zero coefficients.[34] In our hold-out set, we correlate each of these models with both test outcomes and

[33]Though this is a suitable ex-post measure, ex ante this is produced by using L_1 regularization.

[34]We chose this range because the training set contains only $5,188$ tested visits, so we cannot estimate models that use anywhere near the full set of $k = 16,405$ variables.

testing decisions. Two caveats are worth noting. First, we do not assume anything about the model selection properties of LASSO: the particular variables the LASSO chooses is somewhat arbitrary in the setting of correlated, noisy input variables. We are interested only in the complexity of these models, which may be a more stable quantity (Mullainathan and Spiess, 2017). Second, we can only focus on the variables in our data: so we only test hypotheses related to boundedness on observables, not on the variables physicians may use that are unobservable to us.

Figure 5 visually displays the results of this exercise. On the x-axis is k, the model of complexity. On the y-axis is R^2, a measure of goodness of fit (though our results are not specific to this metric: Appendix Figure A.12 shows similar results with another metric, area under the curve). The gray line shows, at each level of complexity, how well that model predicts risk out of sample. The yellow line shows how well the same model predicts physician testing decisions. For risk, we see that R^2 increases at first, then decreases as additional variables lead to over-fitting. For physician decisions, we see a similar pattern: fit increasing with complexity, before starting to reduce. Importantly, however, the two curves hit their peaks at very different levels: for physicians, the empirical optimum is at 49 variables, while for risk it is at 224 variables. The LASSO model that best predicts actual risk is much more complex than the one that best predicts test decisions.

This figure motivates a statistical test. We define two risk predictors: $\widehat{m}_{\mathsf{simple}}(X_{ij})$ which uses the 49 variables above and $\widehat{m}_{\mathsf{complex}}(X_{ij})$ which uses the 224. We will focus on $[\widehat{m}_{\mathsf{complex}}(X_{ij}) - \widehat{m}_{\mathsf{simple}}(X_{ij})]$, the additional risk information provided by the complex model, which we will call "complex risk." We then estimate:

$$T_{ij} = \beta_0 + \beta_1 \widehat{m}_{\mathsf{simple}}(X_{ij}) + \beta_2 [\widehat{m}_{\mathsf{complex}}(X_{ij}) - \widehat{m}_{\mathsf{simple}}(X_{ij})] + \epsilon_{ij} \tag{2}$$

$$Y_{ij} = \gamma_0 + \gamma_1 \widehat{m}_{\mathsf{simple}}(X_{ij}) + \gamma_2 [\widehat{m}_{\mathsf{complex}}(X_{ij}) - \widehat{m}_{\mathsf{simple}}(X_{ij})] + \epsilon_{ij}. \tag{3}$$

If physicians rely only on the simple model, we expect two things to be true. First, $\beta_2 = 0$: the complex risk should not predict testing decisions. Second, $\gamma_2 > 0$: the complex risk should predict yield.

The results are shown in Table 5. Columns (1) and (3) show how the simple risk model predicts both test and yield alone. Columns (2) and (4) then add complex risk. Column (2) shows that as expected, conditional on the simple model, complex risk is not predictive of testing—the coefficient is both very small and statistically insignificant. Column (4) shows that, in contrast, complex risk is predictive of yield and is highly significant. These results provide suggestive evidence that physicians do in fact rely on too simple a model of risk.[35]

How well do physicians use the variables that enter their simpler model? Figure 6 shows the correlation for each of the 49 variables in $\widehat{m}_{\mathsf{simple}}(X_{ij})$ with both test outcome (x-axis) and the test decision (y-axis).[36] We see a strongly positive relationship (R^2: 0.433). While far from proof of the rationality in bounded rationality, this does suggest that physicians do (mostly) correctly weight the variables they do use.

To assess how important boundedness is in explaining under- and over-testing, we look at how much riskier (or less risky) a patient appears if only simple risk is accounted for, which we measure with $\widehat{m}_{\mathsf{simple}}(X_{ij}) - (\widehat{m}(X_{ij})$. We look at its distribution for both low-risk tested patients (the 'over-tested') and high-risk untested patients ('the under-tested'). As shown in Appendix Figure A.13, a full 35.5% of the over-tested come from the top quintile, meaning their simple risk is much larger than their actual risk (compared to 14.5% in the lowest quintile). Likewise, among the undertested, 74.2% come from the bottom quintile, meaning their simple risk is much

[35] Appendix Figure A.12 shows similar results with decision-tree models of risk rather than LASSO models, as well as showing the same result in the nationally representative Medicare claims data.

[36] We standardize test, yield, and predictor variables, and run test and yield on predictors via univariate regressions. So each regression coefficient gives us the correlation and its standard error.

smaller than their actual risk (compared to 7.4% in the top quintile). Boundedness thus appears to be quantitatively important as well for mis-prediction. Physicians identify several good risk predictors that they use if not perfectly, at least modestly well; at the same time, they neglect many other variables that, while individually small, together provide much explanatory power.

Our evidence on boundedness deviates from the traditional perspective of Dawes, Faust, and Meehl (1989), who suggest that people use too complex a model, and that a statistical model would do better by being simpler. In contrast, we find physicians use too simple a model, and a statistical model does better by being more complex. The difference may arise because modern statistical tools better fit complex natural phenomena, echoing recent findings that sparse models fit economic phenomena poorly—despite the appeal (to humans) of 'betting on sparsity' (Gabaix, 2014; Giannone, Lenza, and Primiceri, 2021). In both cases, the underlying reality is complicated, while human judgments are simple.

5.2 Biases in Physician Judgments

Next, we investigate whether physicians over- or under-weight specific variables. One suggestive example is already in Figure 6 where 'Reason for visit: chest pain' is a clear outlier. While a complaint of chest pain does correlate with risk, it correlates even more with testing, suggesting that those with chest pain may be tested at rates above and beyond what is justified by their heightened risk.[37]

The chest pain example suggests two broader hypotheses for why an input might be over-weighted. First, it is highly salient (Tversky and Kahneman, 1974; Bordalo,

[37] Conditional on predicted risk, patients with chest pain are 16 percentage points (578%) more likely to be tested. Appendix Table A.15 shows that for the 10 most common symptoms, nine significantly predict testing after conditioning on predicted risk, including chest pain and shortness of breath (large and positive), and several other smaller negative predictors (e.g., abdominal pain).

Gennaioli, and Shleifer, 2012). Second, it is highly representative of blockage: it is a (the?) stereotypical symptom, both in textbooks and in public understanding (Bordalo et al., 2016). This motivates our exploration of bias: we ask whether variables that are either salient or representative are generally over-weighted.

We study each of these hypotheses in turn, using a similar empirical approach. To assess whether physicians are biased in their use of some subset of variables $\mathcal{W}$, we first create a new risk predictor which uses only those $\mathcal{W}$ variables. Except for the restriction on input variables, this estimator, $\widehat{m}(\mathcal{W}_{ij})$, is built in the training set exactly the same as the original risk predictor. In the holdout set, we then regress yield on full risk (the usual algorithmic measure of risk) as well as this limited risk model $\widehat{m}(\mathcal{W}_{ij})$, analogous to Equation 2 above.[38] We do this to verify that, as expected, conditional on full risk, $\widehat{m}(\mathcal{W}_{ij})$ does not provide additional information. Finally, as our test of whether $\mathcal{W}$ is misused, we regress the test decision T_{ij} again on both full risk $\widehat{m}(X_{ij})$ as well as $\widehat{m}(\mathcal{W}_{ij})$. If physicians over-weight the variables in $\mathcal{W}$ then the coefficient on $\widehat{m}(\mathcal{W}_{ij})$ should positive; if they under-weight, it should be negative.[39]

5.2.1 Symptom Salience

We implement this procedure first for salience. We begin by forming a risk predictor using only symptoms: the most immediate thing the physician sees about a patient, and often stressed in medical education and vignettes. Column (1) of Table 6 shows the results of regressing testing on the full risk predictor; Column (2) then adds the

[38] All regressions control for a vector of risk bins, as well as linear risk, to account for non-linearity of risk in predicted risk. We show the linear coefficient but omit the others for simplicity.

[39] In this exercise, by 'risk' we mean predicted risk. So a bias occurs when an observed variable predicts physician deviations from algorithmic predictions; as the focus is on observed variables, we are less prone to confounding. But still, given the potential for complex relationship between observed and unobserved variables, these results must be taken as suggestive.

new symptom-only risk predictor.[40] We see here that the risk from symptoms is *additionally* predictive of testing, suggesting symptoms as a category are over-weighted.[41]

We then expand this exercise to the entire universe of inputs. We form a set of risk predictors, one for each subset of variables, grouped into the following bins: demographics; prior diagnoses; past procedures done on the patient; prior labs; and prior vital signs. The bins are formed to reflect coherent types of inputs physicians may treat differently. For example, because medical case reports and pedagogy use a standard structure, stressing age, sex, and symptoms (e.g., "A 43 year old man with chest pain," as in the *NEJM*'s Case Records), we conjectured that demographics and symptoms would be highly salient and thus over-weighted. By contrast, the complex, quantitative time series contained in past laboratory studies and vital signs are harder to process and likely less salient. Finally, while some prior diagnoses and procedures relevant to blockages may be salient (e.g., diabetes, prior blockage or stenting procedure), these categories are far broader, including hundreds of other types of information that we also expect to be less salient.

Column (3) shows how these risk predictors correlate with the testing decision. Even after including risk from all other variable subsets, risk from symptoms stays positive (i.e., over-weighted), as is risk from demographic information: a patient in the top quartile of symptom risk is 5.26 percentage points more likely to be tested, relative to other patients, and 0.78 p.p. for demographic risk.[42] This is equivalent to a patient moving from the 50^{th} percentile of true (full) risk to the 89^{th} and 62^{nd} percentile, respectively. Prior quantitative information from laboratory studies and

[40]For space we have left out the yield regressions. These are in Appendix Table A.18 and verify that the symptom-only risk predictor does not predict yield, conditional on full risk.

[41]Abaluck et al. (2016), while they lacked data on symptoms at the visit itself, found that patients with past symptom-based diagnoses were over-tested, consistent with a similar bias.

[42]Appendix Table A.14 further investigates patient demographics, and finds small but significant relationships of specific demographic factors with testing: older patients and women appear to be tested more than their risk merits, while self-reported Hispanic patients are under-tested.

vital signs, though, has a negative sign, suggesting physicians under-weight or neglect this information. Finally, diagnoses are slightly over-weighted while procedures are slightly under-weighted. Taken together, these results are generally supportive of the salience model: clearly salient pieces of data—demographics and symptoms—are attended to more than they should be, while more complex, less salient information—past quantitative vital signs and labs—are neglected.

5.2.2 Representativeness

We use the same method to explore representativeness (Tversky and Kahneman, 1974), as formalized in the model of stereotyping of Bordalo et al. (2016). This predicts that in estimating the probability of blockage for a patient with symptom M, physicians will not use $Pr(B = 1|M = 1)$. Instead they will estimate

$$Pr(B = 1|M = 1) \times g\left(\frac{Pr(M = 1|B = 1)}{Pr(M = 1|B = 0)}\right),$$

where $g(\cdot)$ is monotone. Symptoms more common in patients with blockage, relative to others, will be weighted more heavily than they ought to be.

This model has a crisp empirical prediction: at the same predicted risk, patients with more (less) representative symptoms are more (less) likely to be tested. We investigate this by first identifying the set of symptoms that are potentially representative of blockage. To make this list, we identify those tested patients ultimately found to have blockages, and look back at their presenting symptom (limiting to 16 symptoms with frequency over 0.5% in this population; see Appendix Table A.16). For each symptom M, we calculate its representativeness for blockage: $\frac{Pr(M=1|B=1)}{Pr(M=1|B=0)}$. Nine symptoms have a ratio over 1, which we consider representative of blockage. Some are very common in the general population (e.g., chest pain, shortness of breath) and others are quite rare (e.g., presenting to the ER after a referral by another physician

who is concerned for blockage, or because they were found unresponsive or in cardiac arrest by paramedics). The remaining seven symptoms are more common in the general population than in those with blockage (e.g., dizziness, nausea).

This allows us to build yet another risk predictor, restricting to *representative* symptoms. Column (4) of Table 6 shows the results of adding this to the regression we described previously (Column 2) with the full symptom-based predictor. Adding the representative symptoms makes the full symptom-based predictor small and insignificant. And the coefficient on the representative symptom-based predictor in Column (4) is nearly double the magnitude of the full symptom-based predictor in Column (3).[43] This argues that, while symptoms as a whole may be salient, a small number of representative symptoms push physicians to test far more: they effectively cue the physician's mind to consider blockage. This effect is quantitatively large: the 7% in the highest bin of representative symptom risk are 16.2 p.p. more likely to be tested, corresponding to an increase from the 50^{th} to the 98^{th} percentile of true risk.

Further, as shown in Appendix Figure A.14, patients whose risk comes disproportionately from representative symptoms (i.e., large $\widehat{m}_{\text{represent}}(X_{ij})) - (\widehat{m}(X_{ij}))$ are overrepresented in testing errors. Those in the top quintile of representativeness risk (relative to true risk) make up 34.3% come of the low-risk tested; while the bottom quintile makes up 99.4% of the high-risk untested.[44]

5.3 Implications for Incentive Policies

The simultaneous presence of over- and under-use suggests that simple views of health care like 'less is more' or 'more is more' are insufficiently nuanced. Our results thus add to the growing body of work in health economics arguing for richer

[43] Appendix Table A.18 confirms this new predictor has no incremental value for predicting yield.

[44] An important caveat is that the representative risk is built only on nine indicator variables and thus does not have a wide range, so we view these results as limited.

models of physician behavior (Abaluck et al., 2016; Chan, Gentzkow, and Yu, 2019; Chandra and Staiger, 2020; Kolstad, 2013). Policy makers have long viewed health care through the lens of misaligned incentives that make physicians too eager to test. Implicit in this model is that physicians estimate risk correctly, but simply set too low a threshold. This 'less is more' model, which suggests that high-testing providers are wasteful relative to low-testing ones, has a clear practical implication that drives much of health policy in the US and internationally: create incentives to test less, for example, via reimbursement schemes or capacity constraints. Yet, our finding of systematic biases by physicians calls this approach into question: if physicians mis-predict risk, incentives to cut care may do harm as well as good.

We empirically examine these potentially perverse consequences by asking, when physicians test less, which tests do they cut? The view of traditional models—and the hope of health policy—is that they cut the low-value tests. The top panel of Figure 7 shows this is not the case. Here we graph the probability of testing against predicted risk separately for each of the testing quartiles. Low-testing shifts do cut back on low-value tests: the lowest-risk patients are tested only 0.4% of the time, vs. 3.0% on the highest-testing shifts. But they also cut back on high-value tests: the highest-risk patients are tested 5.8% of the time, vs. 32.3% on the highest-testing shifts. In an absolute sense, high-value tests suffer the biggest decline: 26.5% fewer in low- vs. high-testing regimes; in a relative sense low-value tests fall slightly more than high-value tests (87% vs. 82%). In other words, less testing means less testing for everyone, regardless of risk. The bottom panel replicates these results in our nationally-representative Medicare sample, where we sort hospitals into quintiles based on their testing rate, and again graph testing against predicted risk for each quintile. We see the same result: hospitals that test more test everyone more.[45]

[45]This exercise is inspired by a large health policy literature that makes cross-sectional comparisons

These data provide a reminder that reducing care leads to reductions in care that are *perceived* to be low-value. But when there are prediction errors, what is perceived to be low-value might in fact be extremely valuable. The problem is analogous to 'behavioral hazard' in patients' decision making, where copays lead patients to cut back on both low- and high-value care (Newhouse and Group, 1993; Chandra, Gruber, and McKnight, 2010; Baicker, Mullainathan, and Schwartzstein, 2015; Brot-Goldberg et al., 2017; Handel and Kolstad, 2015; Chandra, Flack, and Obermeyer, 2021). Incentives to reduce care may have perverse consequences.

5.4 The Role of Physician Experience

If incentives do not reduce error, what does? A natural candidate is experience, which provides an opportunity to learn. Though we cannot causally identify the effect of experience, correlations can at least be suggestive. In particular, we study how the correlation between physician decisions and patient risk varies with physician experience (as measured by years since residency). In Table 6, we regress testing on predicted risk, a linear term for experience and an interaction term between experience and risk. Column (5) shows that more experienced physicians test less on average: 1.68 p.p. or 0.05% for every year since residency. At the same time, experienced physicians are better able to match testing decisions to risk: with every year of experience, the lowest-risk patients are 0.04 p.p. (2.81%) less likely to be tested, and the highest-risk 0.58 p.p. (1.06%) are more likely to be tested.[46] These corre-

across hospitals. Naturally, these comparisons can be confounded. While we lack the data to replicate the shift variation experiment, we do have an (albeit weaker) alternative, described in Appendix 8.3. Testing typically requires an overnight stay after ED visits, but since hospital staffing is limited on weekends, patients who come in the day before a weekend are tested less. Figure A.10 shows again that reductions in testing reduce testing for *all* patients, irrespective of their actual risk.

[46]We do not have experience data available for all physicians, so the sample size in this regression decreases from 61,965 to 55,777. As usual, we verify that experience does not additionally predict the yield of testing in Appendix Table A.18.

lations provide suggestive evidence that physicians may become more accurate with experience.

The results on experience in this Section and the results on high- versus low-testing regimes tell distinct stories. On the one hand, experienced physicians both test less and are more accurate. This matches Chan, Gentzkow, and Yu (2019), who show a negative relationship between skill and testing levels. On the other hand, in Section 5.3, we saw that less testing was uncorrelated with accuracy: testing fell across the risk distribution, including the high-risk. This suggest more care is needed to understand the relationship between testing levels and accuracy. Understanding what leads physicians to be more or less accurate—and how that relates to how much they test—is an important and useful open question.

6 Conclusions

Many people believe machine learning will transform health care. Nearly all of these predictions focus on it as a product. For example, algorithms trained to read x-rays can be bought by health systems and used to substitute for radiologists' time. Our work suggests a very different use: machine learning can be used as a tool to understand physician behavior specifically, and health care more broadly.

We believe a lasting contribution of this tool is to more precisely characterize inefficiency. Current empirical approaches in health policy rely on aggregates: for example, do tests on average yield enough positives to justify their costs (Weinstein et al., 1996; Sanders et al., 2016)? In our data, that metric makes testing appear highly efficient, at only $89,714 per life-year. The granularity of algorithmic predictions better captures underlying inefficiencies, revealing both under- and over-use. This reframes the discussion away from how many people get tested—too many, or too few?—to one

about *who* gets tested. In our preferred policy counterfactual, total testing changes by only 1%, but the composition changes radically: 62% of current tests would be cut; and 61% of new tests added; and tests would go from costing $89,714 to $59,390 per life year. The importance of composition in turn calls into question the central role of incentives in policy. By changing the level of testing alone, they may improve one aspect (over-use) while worsening another (under-use).

But we must be careful in comparing human decisions and algorithmic predictions. As we saw, when physician and algorithm disagree, we could not just assume the algorithm was correct: unobserved variables confound algorithmic predictions. This selection bias pervades machine learning applications in medicine and elsewhere, appearing whenever algorithms are trained on data produced by the human decisions they are meant to influence.[47] Once acknowledged, we see they can be tackled: by developing alternative labels grounded in domain expertise, and via quasi-experimental methods from the causal inference toolkit.

Finally, our findings suggest exploring algorithmic predictions to design interventions. Most obviously, because they are built on readily available EHR data, predictions could be delivered to physicians in real time. Rather than replacing their judgment, they can be combined them with physician private information. At the system level, incentives and reimbursements could be tied to patient predicted risk. Or predictions could be used an educational tool during physician training. We found accuracy improves with experience, but algorithmic predictions might help hasten the learning process: trial and error is a costly way to learn in medicine.

[47] In testing decisions, decisions dictate whom we have data for. Our results highlight the importance of taking the 'selective labels' problem seriously (Kleinberg et al., 2018; Kallus and Zhou, 2018; Rambachan, 2021). For treatment decisions, outcomes are treatment polluted; see (Paxton, Niculescu-Mizil, and Saria, 2013) for a discussion.

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". In: *American Economic Review* 106.12, pp. 3730–64.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb (2019). "Exploring the impact of artificial intelligence: Prediction versus judgment". In: *Information Economics and Policy* 47. Publisher: Elsevier, pp. 1–6.

Al-Lamee, Rasha, David Thompson, Hakim-Moulay Dehbi, Sayan Sen, Kare Tang, John Davies, Thomas Keeble, Michael Mielewczik, Raffi Kaprielian, Iqbal S. Malik, and others (2018). "Percutaneous coronary intervention in stable angina (ORBITA): a double-blind, randomised controlled trial". In: *The Lancet* 391.10115, pp. 31–40.

Amsterdam, Ezra A., Nanette K. Wenger, Ralph G. Brindis, Donald E. Casey, Theodore G. Ganiats, David R. Holmes, Allan S. Jaffe, Hani Jneid, Rosemary F. Kelly, Michael C. Kontos, and others (2014). "2014 AHA/ACC guideline for the management of patients with non–ST-elevation acute coronary syndromes: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines". In: *Journal of the American College of Cardiology* 64.24, e139–e228.

Arrow, Kenneth J. (1963). "Uncertainty and the Welfare Economics of Medical Care". In: *The American Economic Review* 53.5. Publisher: American Economic Association, pp. 941–973.

Backus, Barbra E., A. Jacob Six, Johannes C. Kelder, Thomas P. Mast, Frederieke van den Akker, E. Gijis Mast, Stefan HJ Monnink, Rob M. van Tooren, and Pieter AFM Doevendans (2010). "Chest pain in the emergency room: a multicenter validation of the HEART Score". In: *Critical pathways in cardiology* 9.3, pp. 164–169.

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein (2015). "Behavioral Hazard in Health Insurance". In: *The Quarterly Journal of Economics* 130.4, pp. 1623–1667.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer (2016). "Stereotypes". In: *The Quarterly Journal of Economics* 131.4, pp. 1753–1794.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2012). "Salience Theory of Choice Under Risk". In: *The Quarterly Journal of Economics* 127.3, pp. 1243–1285.

— (2020). "Memory, attention, and choice". In: *The Quarterly Journal of Economics* 135.3, pp. 1399–1442.

Brot-Goldberg, Zarek C., Amitabh Chandra, Benjamin R. Handel, and Jonathan T. Kolstad (2017). "What does a deductible do? the impact of cost-sharing on health care prices, quantities, and spending dynamics". In: *The Quarterly Journal of Economics* 132.3, pp. 1261–1318.

Camerer, Colin (2019). "The Economics of Artificial Intelligence: An Agenda". In: ed. by Ajay Agrawal, Joshua Gans, and Avi Goldfarb. NBER conference report. Chap. Artificial Intelligence and Behavioral Economics, pp. 587–610.

Chan, David C. and Jonathan Gruber (2020). "Provider Discretion and Variation in Resource Allocation: The Case of Triage Decisions". In: *AEA Papers and Proceedings* 110, pp. 279–283.

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

Chandra, Amitabh, Evan Flack, and Ziad Obermeyer (2021). "The health costs of cost-sharing". In: *National Bureau of Economic Research Working Paper 28439*.

Chandra, Amitabh, Jonathan Gruber, and Robin McKnight (2010). "Patient Cost-Sharing and Hospitalization Offsets in the Elderly". In: *American Economic Review* 100.1, pp. 193–213.

Chandra, Amitabh and Douglas O. Staiger (May 2020). "Identifying Sources of Inefficiency in Healthcare". In: *The Quarterly Journal of Economics* 135.2. Publisher: Oxford Academic, pp. 785–843.

Croskerry, Pat (2002). "Achieving quality in clinical decision making: cognitive strategies and detection of bias". In: *Academic Emergency Medicine* 9.11, pp. 1184–1204.

Dawes, R. M., D. Faust, and P. E. Meehl (Mar. 1989). "Clinical versus actuarial judgment". In: *Science (New York, N.Y.)* 243.4899, pp. 1668–1674.

Elstein, Arthur S. (1999). "Heuristics and biases: Selected errors in clinical reasoning". In: *Academic Medicine* 74.7, pp. 791–794.

Fisher, Elliott S., David E. Wennberg, Thérèse A. Stukel, Daniel J. Gottlieb, F. L. Lucas, and Étoile L. Pinder (Feb. 2003). "The Implications of Regional Variations in Medicare Spending. Part 2: Health Outcomes and Satisfaction with Care". In: *Annals of Internal Medicine* 138.4, pp. 288–298.

Gabaix, Xavier (2014). "A sparsity-based model of bounded rationality". In: *The Quarterly Journal of Economics* 129.4, pp. 1661–1710.

— (2019). "Handbook of Behavrioal Economics – Foundations and Applications". In: ed. by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson. Vol. 2. Chap. Behavioral Inattention, pp. 261–343.

Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri (2021). "Economic predictions with big data: The illusion of sparsity". In: Publisher: ECB Working Paper.

Graber, Mark L., Nancy Franklin, and Ruthanna Gordon (July 2005). "Diagnostic Error in Internal Medicine". In: *Archives of Internal Medicine* 165.13, pp. 1493–1499.

Hamon, Martial, Jean-Claude Baron, Fausto Viader, and Michèle Hamon (2008). "Periprocedural stroke and cardiac catheterization". In: *Circulation* 118.6, pp. 678–683.

Handel, Benjamin R. and Jonathan T. Kolstad (2015). "Health insurance for "humans": Information frictions, plan choice, and consumer welfare". In: *American Economic Review* 105.8, pp. 2449–2500.

IOM, (Institute of Medicine) (2015). *Improving Diagnosis in Health Care*. Washington, DC.

Kahneman, Daniel and Amos Tversky (1972). "Subjective probability: A judgment of representativeness". In: *Cognitive psychology* 3.3, pp. 430–454.

Kallus, Nathan and Angela Zhou (Dec. 2018). "Confounding-robust policy improvement". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Red Hook, NY, USA, pp. 9289–9299.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018). "Human decisions and machine predictions". In: *The Quarterly Journal of Economics* 133.1, pp. 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015). "Prediction Policy Problems". In: *American Economic Review* 105.5, pp. 491–95.

Kolstad, Jonathan T. (2013). "Information and quality when motivation is intrinsic: Evidence from surgeon report cards". In: *American Economic Review* 103.7, pp. 2875–2910.

Mahoney, Elizabeth M., Claudine T. Jurkovitz, Haitao Chu, Edmund R. Becker, Steven Culler, Andrzej S. Kosinski, Debbie H. Robertson, Charles Alexander, Soma Nag, John R. Cook, and others (2002). "Cost and cost-effectiveness of an early invasive vs conservative strategy for the treatment of unstable angina and non–ST-segment elevation myocardial infarction". In: *JAMA* 288.15, pp. 1851–1858.

Mullainathan, Sendhil (2002). “A memory-based model of bounded rationality”. In: *The Quarterly Journal of Economics* 117.3, pp. 735–774.

Mullainathan, Sendhil and Jann Spiess (2017). “Machine learning: an applied econometric approach”. In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Newhouse, Joseph P. and Rand Corporation Insurance Experiment Group (1993). *Free for all?: lessons from the RAND health insurance experiment*.

Norris, Samuel (2019). “Examiner inconsistency: Evidence from refugee appeals”. In: *University of Chicago, Becker Friedman Institute Working Paper* 2018-75.

Obermeyer, Ziad, Brent Cohn, Michael Wilson, Anupam B. Jena, and David M. Cutler (Feb. 2017). “Early death after discharge from emergency departments: analysis of national US insurance claims data”. In: *British Medical Journal* 356.

Pauly, Mark V. (1968). “The Economics of Moral Hazard: Comment”. In: *The American Economic Review* 58.3. Publisher: American Economic Association, pp. 531–537.

Paxton, Chris, Alexandru Niculescu-Mizil, and Suchi Saria (2013). “Developing predictive models using electronic medical records: challenges and pitfalls”. In: *AMIA Annual Symposium Proceedings*. Vol. 2013, pp. 1109–1115.

Pope, J. Hector, Tom P. Aufderheide, Robin Ruthazer, Robert H. Woolard, James A. Feldman, Joni R. Beshansky, John L. Griffith, and Harry P. Selker (2000). “Missed diagnoses of acute cardiac ischemia in the emergency department”. In: *New England Journal of Medicine* 342.16, pp. 1163–1170.

Prasad, Vinay, Michael Cheung, and Adam Cifu (2012). “Chest pain in the emergency department: The case against our current practice of routine noninvasive testing”. In: *Arch Intern Med* 172.19, pp. 1506–1509.

Rambachan, Ashesh (2021). “Identifying prediction mistakes in observational data”. In: *Working Paper*.

Redelmeier, Donald A., Lorraine E. Ferris, Jack V. Tu, Janet E. Hux, and Michael J. Schull (Feb. 2001). “Problems for clinical judgement: introducing cognitive psychology as one more basic science”. In: *CMAJ: Canadian Medical Association Journal* 164.3, pp. 358–364.

Sanders, Gillian D., Peter J. Neumann, Anirban Basu, Dan W. Brock, David Feeny, Murray Krahn, Karen M. Kuntz, David O. Meltzer, Douglas K. Owens, Lisa A. Prosser, and others (2016). “Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine”. In: *JAMA* 316.10, pp. 1093–1103.

Shanmugam, Vimalraj Bogana, Richard Harper, Ian Meredith, Yuvaraj Malaiapan, and Peter J Psaltis (Mar. 2015). “An overview of PCI in the very elderly”. In: *Journal of Geriatric Cardiology : JGC* 12.2, pp. 174–184.

Simon, Herbert A. (1955). “A behavioral model of rational choice”. In: *The Quarterly Journal of Economics* 69.1, pp. 99–118.

Sims, Christopher A. (2003). “Implications of rational inattention”. In: *Journal of monetary Economics* 50.3, pp. 665–690.

Tversky, Amos and Daniel Kahneman (1974). “Judgment under uncertainty: Heuristics and biases”. In: *Science* 185.4157, pp. 1124–1131.

Wei, Wei-Qi, Qiping Feng, Peter Weeke, William Bush, Magarya S. Waitara, Otito F. Iwuchukwu, Dan M. Roden, Russell A. Wilke, Charles M Stein, and Joshua C. Denny (Apr. 2014). “Creation and Validation of an EMR-based Algorithm for Identifying Major Adverse Cardiac Events while on Statins”. In: *AMIA Summits on Translational Science Proceedings* 2014, pp. 112–119.

Weinstein, Milton C., Louise B. Russell, Marthe R. Gold, and Joanna E. Siegel, eds. (1996). *Cost-effectiveness in health and medicine*.

Ægisdóttir, Stefanía, Michael J. White, Paul M. Spengler, Alan S. Maugherman, Linda A. Anderson, Robert S. Cook, Cassandra N. Nichols, Georgios K. Lampropoulos, Blain S. Walker, Genna Cohen, and Jeffrey D. Rush (May 2006). “The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction”. In: *The Counseling Psychologist* 34.3, pp. 341–382.

Figures and Tables

Table 1: Summary Statistics: Patient Characteristics

	All	Tested	Untested
Patients (N)	129,859	6,088	123,771
Visits (N)	246,265	7,320	238,945
Demographics			
Age (years)	42	58	42
	(0.033)	(0.146)	(0.033)
Female	0.612	0.459	0.616
	(<0.001)	(0.006)	(<0.001)
Black	0.262	0.216	0.264
	(<0.001)	(0.005)	(<0.001)
Hispanic	0.237	0.145	0.24
	(<0.001)	(0.004)	(<0.001)
White	0.436	0.588	0.431
	(<0.001)	(0.006)	(0.001)
Heart Disease Risk			
Past Heart Disease	0.122	0.393	0.114
	(<0.001)	(0.006)	(<0.001)
Diabetes	0.142	0.294	0.137
	(<0.001)	(0.005)	(<0.001)
Hypertension	0.253	0.517	0.245
	(<0.001)	(0.006)	(<0.001)
Cholesterol	0.163	0.418	0.156
	(<0.001)	(0.006)	(<0.001)
Any Risk Factor	0.361	0.626	0.352
	(<0.001)	(0.006)	(<0.001)
Triage Shifts			
Number of Shifts (N)	3,951		
Patients per Shift (N)	62.3		

Notes: Main sample descriptive statistics (mean (SE)). Numbers are fractions unless otherwise noted. Past heart disease is the fraction with any diagnosis of heart problems (ischemia), stroke, or peripheral vascular disease prior to the visit. Frequency of individual risk factors (diabetes, hypertension, high cholesterol) is shown, along with the fraction with any of these risk factors.

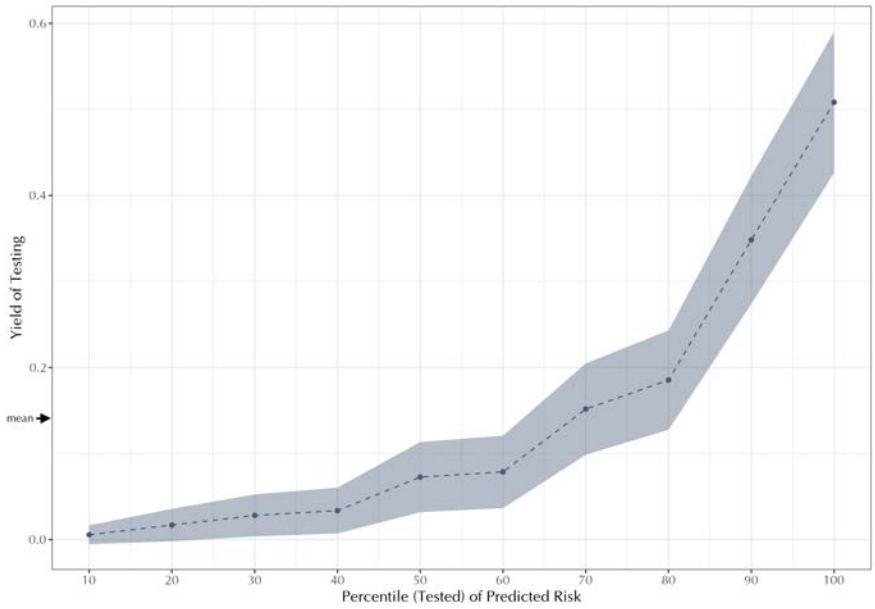
Table 2: Summary Statistics: Physician Choices and Patient Outcomes

	All	Tested	Untested
Tested	0.03	-	-
	(<0.001)	-	-
Catheterization	0.013	-	-
	(<0.001)	-	-
Stress Testing	0.02	-	-
	(<0.001)	-	-
Yield of Testing	0.004	0.146	0
	(<0.001)	(0.004)	(<0.001)
Stenting	0.004	0.129	0
	(<0.001)	(0.004)	(<0.001)
Open-heart Surgery	0.001	0.018	0
	(<0.001)	(0.002)	(<0.001)
Adverse Events	0.019	0.261	0.011
	(<0.001)	(0.005)	(<0.001)
Diagnosed Event	0.016	0.253	0.008
	(<0.001)	(0.005)	(<0.001)
Death	0.004	0.017	0.004
	(<0.001)	(0.002)	(<0.001)
One-Year Mortality	0.016	0.048	0.015
	(<0.001)	(0.002)	(<0.001)
Physician Suspicion			
ECG Done	0.294	1.0	0.275
	(<0.001)	(0.004)	(<0.001)
Troponin Done	0.131	0.792	0.111
	(<0.001)	0.005	(<0.001)
Diagnosed Heart Damage	0.023	0.391	0.012
	(<0.001)	(0.006)	(<0.001)
Positive Troponin	0.025	0.221	0.019
	(<0.001)	(0.005)	(<0.001)
Troponin Result (ng/ml)	0.278	0.72	0.124
	0.003	0.005	0.002

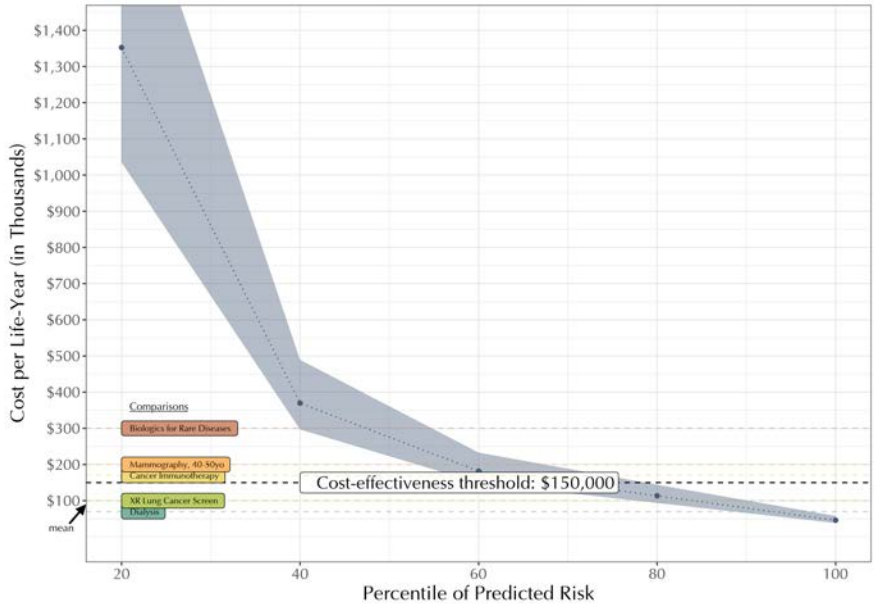
Notes: Main sample outcomes (mean (SE)). Numbers are fractions unless otherwise noted. Test and yield rates use a 10-day window, adverse events a 30-day window. ECG and troponin are low-cost tests done for even a very slight suspicion of blockage. Diagnosed heart damage reflects codes for myocardial infarction or ischemia assigned at the end of a visit. Positive troponin is a laboratory test that indicates damage to heart muscle.

Figure 1: Yield and Cost-Effectiveness of Testing in Tested Patients

(a) Realized Yield of Testing



(b) Cost-Effectiveness of Testing



Notes: Realized yield of testing (top) and cost-effectiveness (bottom) of tests (y-axis; sample mean shown with an arrow) in the tested, by bin of predicted risk (x-axis). Bins are deciles of predicted risk. The cost-effectiveness line shows our preferred specification, and the shaded interval shows sensitivity to a range of estimated treatment effects from the literature. For comparison, we include cost-effectiveness estimates of several other tests and treatments.

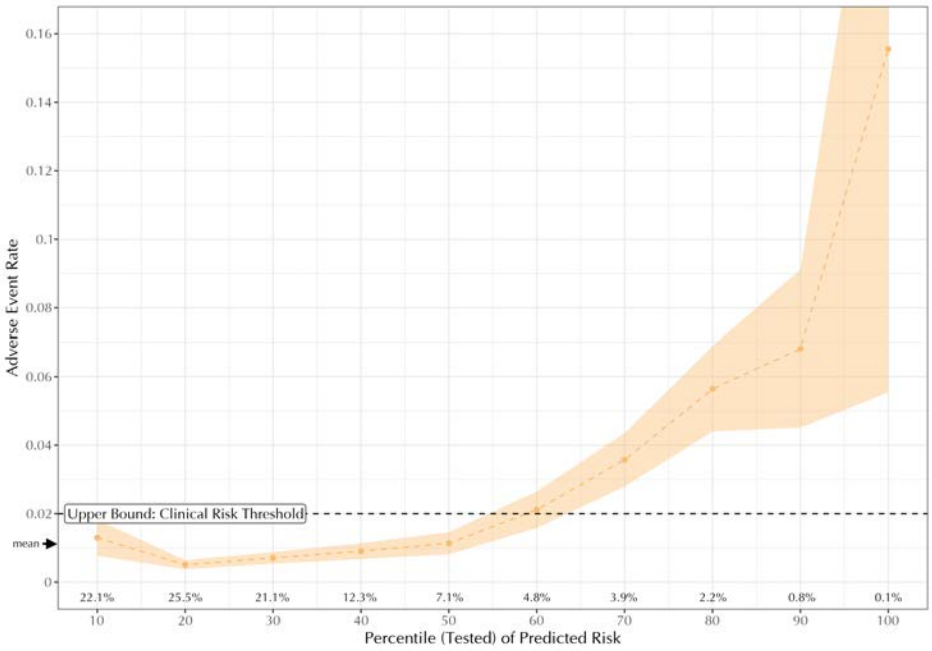
Table 3: Realized Yield, Cost-Effectiveness, and Testing Rate

	Yield Rate (SE) (1)	Cost-Effectiveness ($) (Lower–Upper Bound) (2)	Test Rate (SE) (3)
Full Sample	0.146	89,714	0.03
	(0.004)	(74,152-113,543)	(<0.001)
By Risk Bin			
1	0.011	1,352,466	0.012
	(0.006)	(1,034,814-1,951,515)	(<0.001)
2	0.036	318,603	0.017
	(0.01)	(257,296-418,265)	(0.001)
3	0.07	192,482	0.047
	(0.014)	(157,552-247,314)	(0.002)
4	0.168	114,146	0.088
	(0.02)	(94,154-144,914)	(0.004)
5	0.429	46,017	0.383
	(0.026)	(38,178-57,907)	(0.016)
N	1,784	1,784	61,965

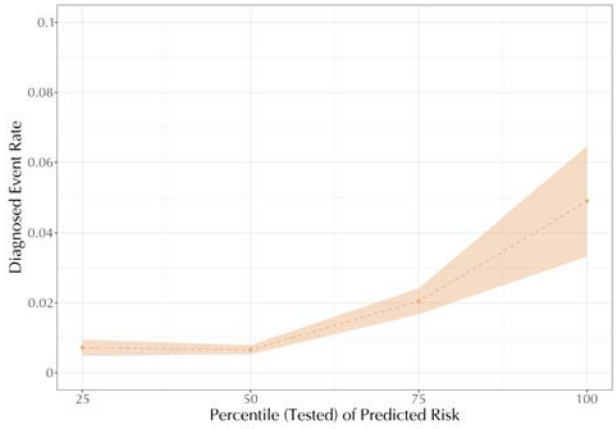
Notes: Yield of testing (1) and cost-effectiveness of testing (2) in the tested, and test rate across all visits (3), by bin of predicted risk. Bins are quintiles of risk, defined in the tested population (so bins are equally sized in Columns (1) and (2), but not in (3)).

Figure 2: Adverse Events in Untested Patients (30 Days After Visits)

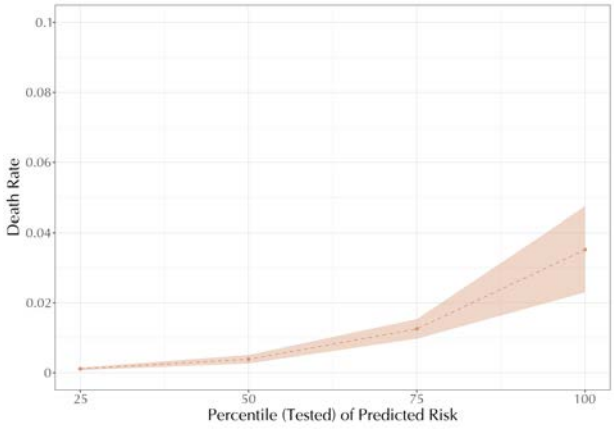
(a) Any Adverse Event



(b) Diagnosed Blockage or Arrythmia



(c) Death

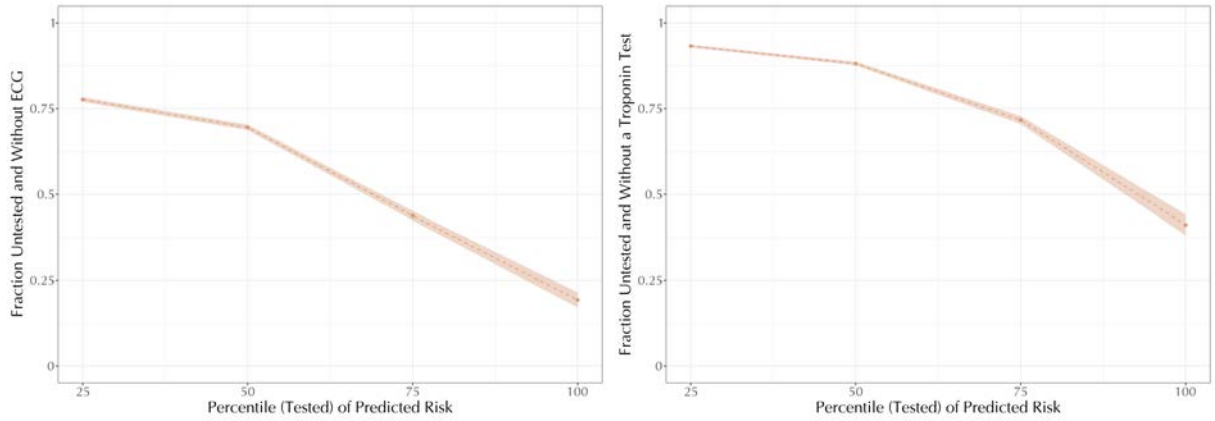


Notes: Rate of adverse events over the 30 days following visits (y-axis) among untested patients, by bin of predicted risk (x-axis). Risk bins are formed as deciles of predicted risk in the tested, for comparison (so bins are not equally sized). Top panel (a) shows the total event rate. The horizontal line shows the 2% threshold above which testing is recommended by clinical guidelines; the highest-risk 14% (top 6 bins) have a rate significantly above 2b. The percent in each bin is shown above the x-axis. Top of the highest 95% CI truncated. Bottom panels disaggregate two categories of adverse events that make up the total: (b) diagnosed adverse events (heart damage, confirmed with laboratory biomarkers; and cardiac arrest) (c) death (via linkage to Social Security data); bins are formed as quartiles of predicted risk (because outcomes are less frequent).

Figure 3: Adverse Events in Untested and Unsuspected Patients (30 Days)

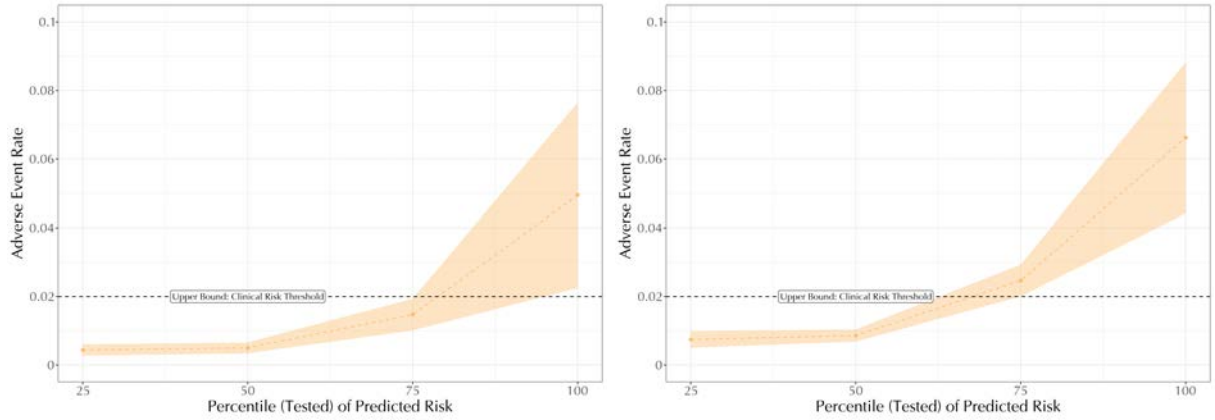
(a) Fraction of Untested, No ECG

(b) Fraction of Untested, No Troponin



(c) Adverse Events, No ECG

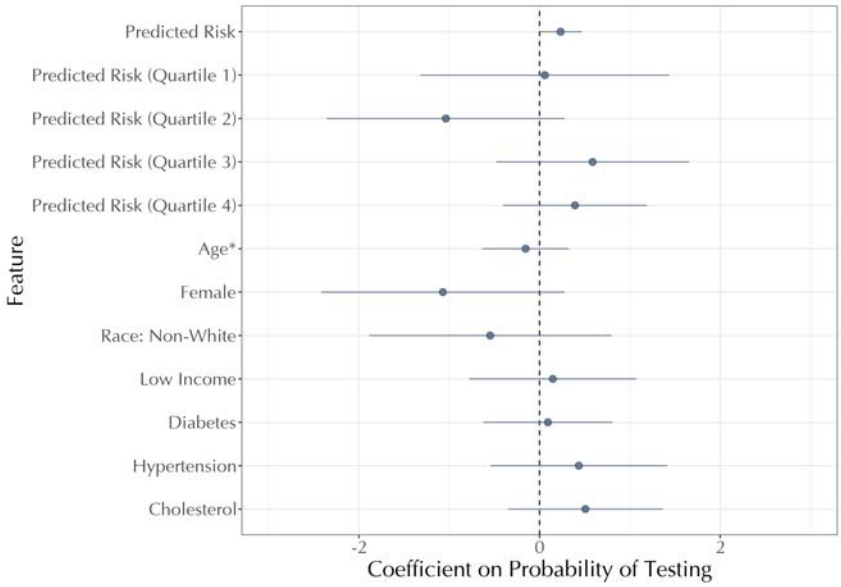
(d) Adverse Events, No Troponin



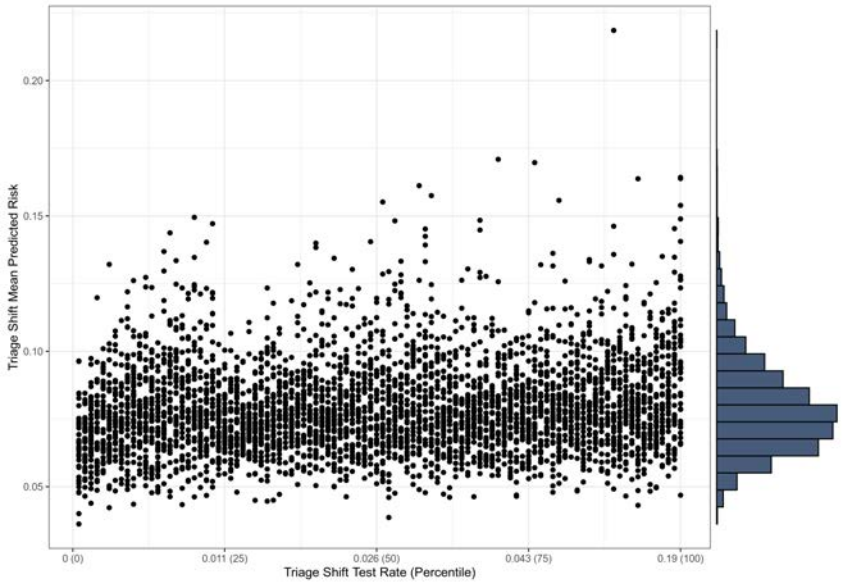
Notes: Top panels: fraction of untested patients in whom physicians do not appear to suspect blockage, based on lack of two low-cost tests done in any case of suspected blockage: (a) an electrocardiogram (ECG), and (b) a troponin test (a laboratory study). Rates are shown by risk bins, which are formed as quartiles of predicted risk in the tested for comparison (so bins are not equally sized). Bottom panels: rate of adverse events (diagnosed events and death after visits (y-axis), by bin of predicted risk (x-axis), among patients lacking (c) an ECG, and (d) a troponin. The horizontal line shows the clinical threshold above which testing is recommended.

Figure 4: Balance on Observables Across Triage Shfits

(a) Variation in Testing Rate and Observables, by Shift Testing Rate



(b) Variation in Average Predicted Risk, by Shift Testing Rate



Notes: Panel (a) shows results of balance checks in a 'natural experiment,' in which patients arriving during different triage shifts are tested at higher or lower rates. Each row corresponds to a regression of a pre-triage variable on leave-one-out shift testing rate. Each point shows the coefficient and confidence interval on leave-one-out shift testing rate. Panel (b) plots, for each shift, the average testing rate for all patients who arrive in that shift (in percentile terms, x-axis) and the average predicted risk of those patients (y-axis). Each point represents one of 3,951 shifts in our dataset, and the density plot on the right shows overall distribution of mean risk. *Age is divided by 100 for scale.

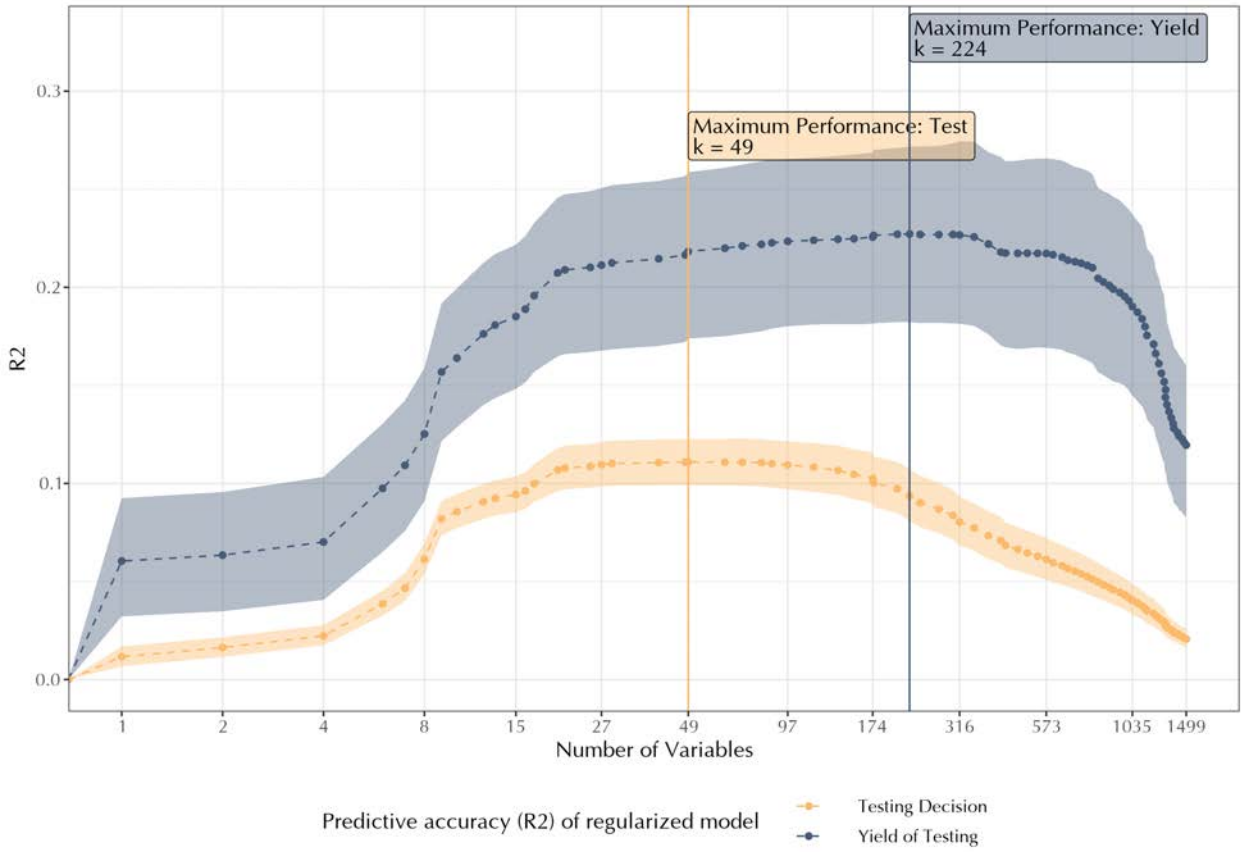
Table 4: Marginal Effect of Increasing Testing, Using Shift Testing Variation

	Diagnosed Event (31-365)	Death (31-365)	Death (0-365)
	(1)	(2)	(3)
Average Effect			
Predicted Risk	0.05***	0.15***	0.25***
	(0.005)	(0.01)	(0.01)
Shift Test Rate	0.02	0.005	0.005
	(0.01)	(0.01)	(0.02)
Observations	123,289	123,289	123,289
Heterogeneous Effect By Risk			
Predicted Risk	0.06***	0.17***	0.27***
	(0.01)	(0.01)	(0.01)
Shift Test Rate	0.04**	0.04**	0.04*
	(0.02)	(0.02)	(0.02)
Predicted Risk	−0.25*	−0.49***	−0.43**
× Shift Test Rate	(0.15)	(0.17)	(0.20)
Observations	123,289	123,289	123,289
Outcome Rate	0.018	0.012	0.016
Outcome Rate, Top Risk Bin	0.027	0.046	0.077

*p<0.1; **p<0.05; ***p<0.01

Notes: Top panel: Regression of diagnosed adverse events (Column 1) and death over days 31–365 after visits (Column 2) on leave-one-out shift testing rate. We use 31–365 days because tested patients are mechanically more likely to be diagnosed with heart problems than untested patients in the first 30 days. Our mortality data do not suffer from this difference in ascertainment, so death over the full year after visits is also shown (Column 3). Bottom panel: The same regression, but with an additional interaction term that allows the effect of testing to vary by predicted risk. Outcome rates, overall and in the top risk quintile, are shown below. Controls for time fixed effects (year, week of year, day of week, and hour of day) and patient risk are included but not shown.

Figure 5: Explanatory Power of Simple vs. Complex Models of Risk



Notes: Using a LASSO model of predicted risk (part of our full ensemble risk model), we preserve all risk models along the regularization path for $k \in [0, 1500]$: the best fit linear model that uses at most k non-zero coefficients. We then measure the explanatory power of these models for physician testing decisions, and for patient risk (measured by yield of testing). The x-axis shows k, the number of variables retained as the regularization penalty is decreased, moving from left to right (we do not show the full path, out to $k = 16,381$, for computational reasons). The y-axis shows R^2 for testing decisions (gray line), and patient risk (yellow line). Uncertainty is shown in the shaded intervals, calculated by bootstrapping. The vertical lines show the complexity of the model that explains the most variance in physician decisions (k_h^*) and risk (k_r^*).

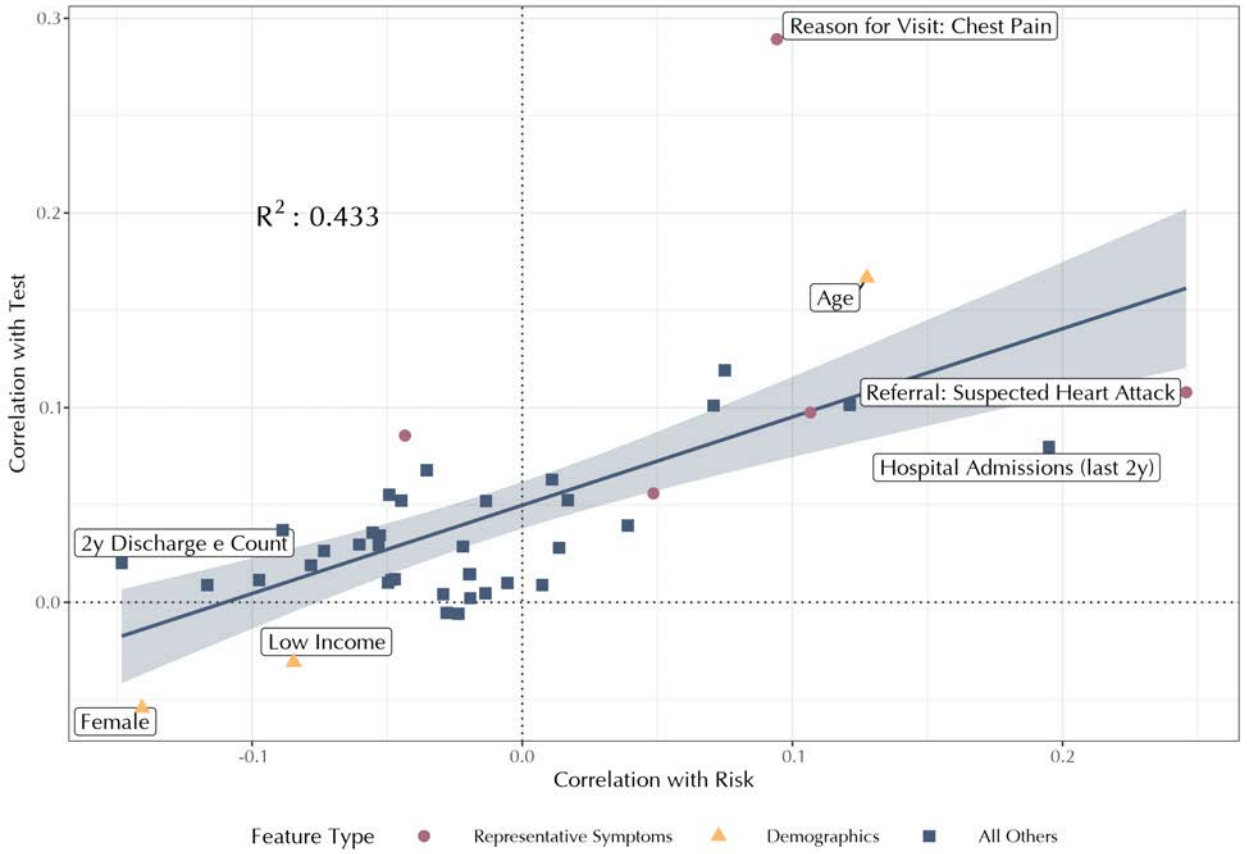
Table 5: Evidence for Physician Boundedness

	Test		Yield	
	(1)	(2)	(3)	(4)
Predicted Risk, Simple	1.357^{***}	1.358^{***}	1.528^{***}	1.319^{***}
($k = 49$)	(0.015)	(0.016)	(0.068)	(0.081)
Incremental Risk, Complex		-0.005		1.099^{***}
($k = 224$)		(0.033)		(0.236)
Constant	-0.059^{***}	-0.059^{***}	-0.076^{***}	-0.043^{***}
	(0.001)	(0.001)	(0.012)	(0.014)
Observations	61,821	61,821	1,834	1,834
R^2	0.111	0.111	0.218	0.227

$^{*}p < .1, ^{**}p < .05, ^{***}p < .01$

Notes: Tests of the explanatory power of two versions of predicted risk, for physician testing decisions and patient risk (yield of testing). We first identify the simple risk model of complexity $k_h^* = 49$ that explains the most variance in physician decisions (here: Predicted Risk, Simple). We then subtract this prediction from the risk model of complexity $k_h^* = 224$ that explains the most variance in patient risk (here: Incremental Risk, Complex). Columns (1) and (3) show how the simple risk model predicts both test and yield alone. Columns (2) and (4) then add the complex model's incremental contribution to predicted risk.

Figure 6: Simple Risk Variables: Correlation with Testing and Predicted Risk



Notes: For the simple risk model of complexity $k_h^* = 49$ that best predicts physicians' testing decisions, we show univariate correlations of each included variable with the physician's testing decision (y-axis) and patient risk (x-axis). Each point is one of the 49 included variables, with separate shapes denoting different categories of inputs. Some outlier points of interest are labeled.

Table 6: Symptom Salience and Representativeness

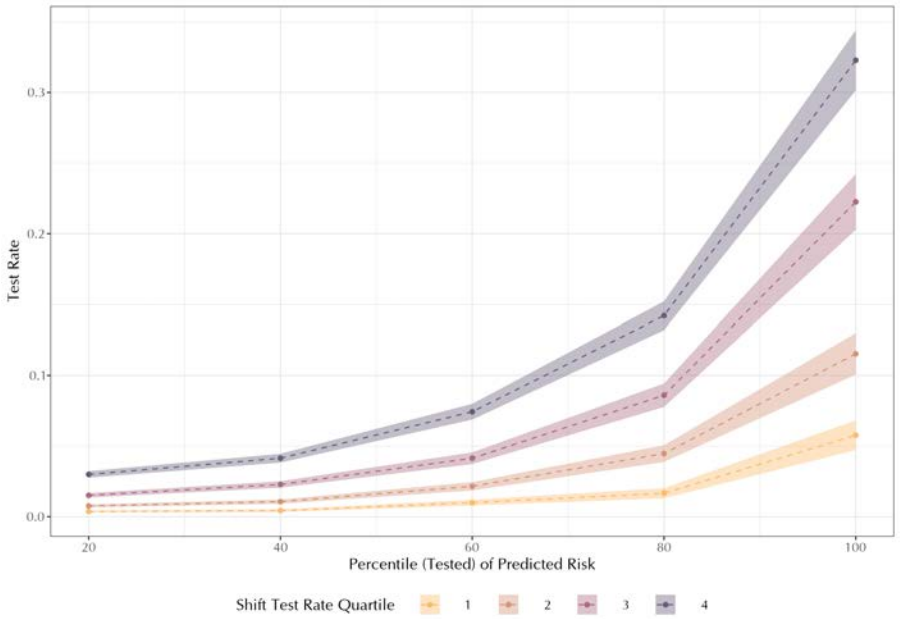
	Test				
	(1)	(2)	(3)	(4)	(5)
Predicted Risk, Full	0.872***	0.715***	0.756***	0.619***	0.755***
	(0.053)	(0.049)	(0.061)	(0.045)	(0.066)
Predicted Risk, Subsets					
All Symptoms		0.888***	0.860***	0.273***	
		(0.052)	(0.057)	(0.061)	
Representative Symptoms				1.283***	
				(0.121)	
Demographics			0.139***		
			(0.031)		
Prior Diagnoses			0.046**		
			(0.021)		
Prior Procedures			−0.053*		
			(0.030)		
Prior Lab Results and Vital Signs			−0.209***		
			(0.019)		
Physician Experience					
Experience (years)					−0.0005**
					(< 0.001)
Experience × Risk					0.011***
					(0.005)
Constant	−0.014***	−0.099***	−0.081***	−0.171***	
	(0.002)	(0.005)	(0.008)	(0.010)	
Observations	61,938	61,938	61,938	61,938	55,777
R^2	0.084	0.106	0.113	0.118	0.082

$^{*}p < .1, ^{**}p < .05, ^{***}p < .01$

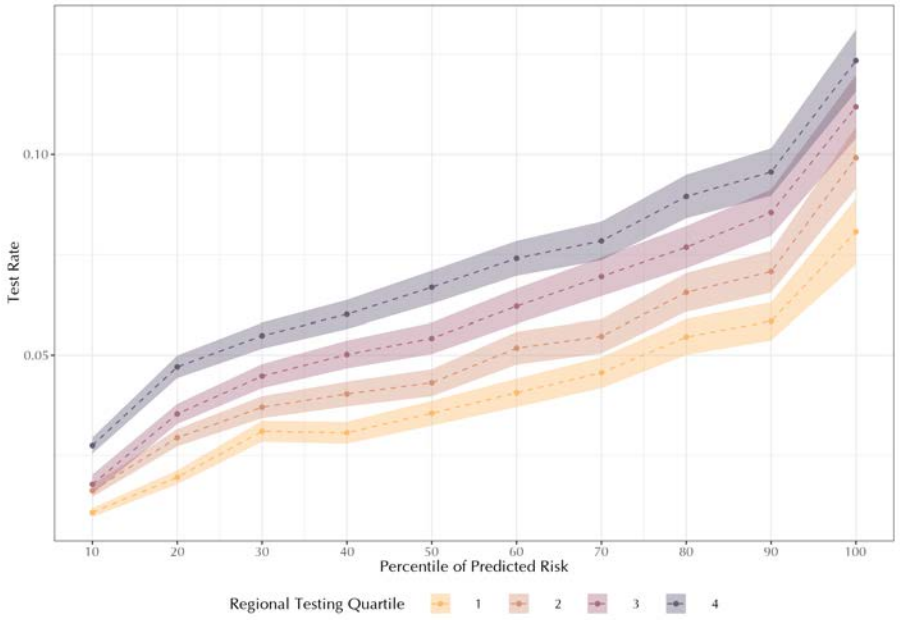
Notes: Column (1) regresses testing on predicted risk. Column (2) adds a risk predictor formed using only symptom inputs. Column (3) adds risk predictors to (2), formed using other input categories. Column (4) adds another risk predictor to (2), formed from only nine *representative* symptoms. Column (5) regresses testing on predicted risk and physician experience (linear and interacted with risk). All models control for non-linear risk terms (not shown). Appendix Table A.18 shows none of these variables besides full predicted risk predict yield of testing.

Figure 7: Variation in Testing Rates by Predicted Risk

(a) Hospital Sample



(b) National Medicare Sample



Notes: Panel (a) shows variation in testing rates by predicted risk, in our 'natural experiment' where patients are tested at higher or lower rates based on the triage team working when they arrive. Panel (b) shows variation in testing rate by predicted risk, across all hospitals in the US. Hospitals are binned into quartiles based on the overall testing rate of the hospital referral region in which they are located, to mirror cross-sectional analyses in the literature.