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 March 2021

Previously circulated as "Who is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error" and "A Machine Learning Approach to Low-Value Health Care: Wasted Tests, Missed Heart Attacks and Mis-Predictions." Authors contributed equally. We acknowledge grants from the National Institutes of Health (DP5OD012161, P01AG005842) and the Pershing 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 outstanding 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 March 2021
JEL No. C55,D8,D84,D9,I1,I13

ABSTRACT

A disease cannot be treated if it is not diagnosed. We study the physician's decision of when and whom to test, a key part of the diagnostic process. To do so, we create algorithmic predictions of a test's outcome for each patient, then contrast these with the physician's testing decision. Applying this approach to data on testing for heart attacks (acute coronary syndromes) in the ER reveals three findings. First, physicians frequently order tests that the algorithm predicts will be low-yield. Data from a hold-out set, which the algorithm has never seen, confirms they are indeed low-yield. At a threshold of $150,000 per life-year, 49% of all tests should be cut. By contrast, testing on average appears cost-effective ($86,683 per life-year), illustrating how algorithmic predictions provide a more granular way to identify low-value health care than standard approaches, which do not account for patient differences. Second, physicians also fail to test many patients that the algorithm predicts will be high-yield. But these choices need not be mistakes: we cannot validate the algorithm's predictions since the untested do not have test results. To test for error, we look to the health outcomes of high-risk untested patients. In the 30 days immediately after their visits, they go on to have adverse cardiac events, including death, at rates high enough that clinical guidelines suggest they should have been tested. This suggests they were indeed high- risk. Using natural variation in testing rates (driven by some triage teams testing more than others), we also show that increased testing greatly reduces adverse events—but only in high-risk patients. Finally, we investigate the behavioral underpinning of errors. Physicians appear to be boundedly rational: they rely on a simpler model of risk than best fits the data. They also exhibit systematic biases: they overweight salient symptoms; and in particular the ones that are stereotypical of heart attack. Together these results suggest a central role for physician error in generating low-value care, and suggest large returns to policy solutions that address both over- and under-use.

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 at Berkeley
2121 Berkeley Way
Berkeley, CA 94704
zobermeyer@berkeley.edu

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

1 Introduction

The quality of health care depends not only on the quality of treatment decisions, but also on the quality of diagnostic judgments. Diagnosis, however, is not easy. Physicians must integrate diverse information—results from a physical exam, the patient's reported symptoms, their past history, etc.—into a probabilistic judgment about what, if anything, is wrong with the patient. Since many of these judgments take place inside the physician's mind, studying diagnosis is also not easy. One key element of the process, though, can be measured: which tests are ordered and what the results are. Since tests are the primary means of validating suspicions, these decisions provide an empirical window into the quality and nature of diagnostic judgments.

We suggest that such testing data can be analyzed profitably using machine learning techniques. In deciding whom to test, the physician must in effect form a prediction. A test sure to come back negative is a waste; and except at the extreme, the value of a test increases with the probability it will come back positive.[1] As such, efficient testing is grounded in effective predicting. Decisions such as these, based on predictions, are ideal for applying machine learning, because algorithms provide explicit predictions against which actual decisions can be contrasted (Kleinberg, Ludwig, Mullainathan, and Obermeyer, 2015; Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan, 2017). The highly tailored nature of these predictions can provide insights at a new and granular level. In contrast, we typically calculate the cost-effectiveness of a test using average test yield across all patients, masking the heterogeneity in what are ultimately patient-by-patient decisions. By mapping out the full distribution of cost-effectiveness, comparisons with algorithmic predictions can better illuminate decision-making.

We apply this approach to the case of testing for heart attack—more precisely, 'acute coronary syndromes,' new blockages in the coronary arteries supplying the heart. Patients with heart attack need treatment to prevent complications: untreated blockages can cause heart failure and chronic pain (angina), or arrythmia and sudden death.[2] Blockages cannot be diagnosed from symptoms alone. For example, the most common symptom, chest pain, could also be the result of a pinched nerve, or acid reflux, for example. And the laboratory tests and electrocardiograms done in the ER can be suggestive of blockage, but not conclusive. As such, physicians rely on an invasive procedure, cardiac catheterization, which they can perform either directly or after a lower-cost 'stress test.' Since these tests are costly, they are applied only only when the probability of a heart attack is judged to be sufficiently high. Predicting these probabilities directly thus allows us to assess whether the right patients are

[1] We assume a positive test is one that changes the (default) treatment decision; that the information value of the test is summarized by its effect on treatment decisions; and that downstream treatments are effective and useful, with homogeneous benefits. We discuss these assumptions, and the possibility of heterogeneous benefits, below. We also assume that the ex ante probability of a positive test is far from 1 such that testing still adds information.

[2] These treatments have large and incontrovertible benefits in the emergency setting, which we study here. They are less clearly useful in 'stable' coronary artery disease, i.e., patients with longstanding symptoms presenting to places besides the ER; see Al-Lamee et al. (2018).

tested. We view testing for heart attack as both an important problem in its own right, and as a 'model system' for applying machine learning to study diagnostic judgments more generally.

We implement our approach on electronic health record data from a large academic medical center. These data span 2010-15 and for each patient visit, contain information on diagnoses, procedures, treatments, physician notes as well as lab tests and values. We focus on visits by generally healthy patients—free of cancer, dementia, etc., living independently, not receiving end-of-life care—which along with other exclusions, leaves us with 246,874 visits. A subset of these is used to train an ensemble machine learning model, that uses 16,831 variables to predict whether a given test will reveal a coronary blockage (measured by whether the blockage was treated). We refer to this as predicted risk. In addition, we build a cost effectiveness model that, for a set of patients at a given predicted risk, calculates the implied cost of a life year for testing that set, based on known benefits and costs of treatment. To check the generality of our results, we replicate them in a 20% sample of 20,059,154 emergency visits by Medicare beneficiaries from 2009-2013. These data, based on insurance claims, are less detailed. But because they are nationally representative, allow us to explore the relevance of our results for health policy.

We first examine those patients whom physicians choose to test. Our strategy is to use the algorithm's risk predictions to identify *potentially* low-risk patients, in whom testing might not be useful. Importantly, we do not simply assume the algorithm's predictions to be accurate. Instead, we look at *realized* test results to see who was right—the algorithm or the physician. This allows us to calculate the actual value of testing ex post, as a function of predicted risk ex ante, and identify predictably low-value testing. The value of testing is typically adjudicated on the basis of its average yield, and in our setting, the average test has an implied cost effectiveness of $86,683 per life year. At typical US life-year valuations of $100-150,000 (Neumann, Cohen, and Weinstein, 2014), this would be considered cost-effective. But this aggregate statistic hides a great deal of highly inefficient testing: binning patients by predicted risk reveals that, at a threshold of $150,000 per life year, 49% of tests should be cut based on predicted risk. These tests in the bottom half of the risk distribution cost $357,787 per life year, vs. a highly cost-effective $62,733 in the top half. Our results illustrate the stark difference between average cost-effectiveness and its full distribution, as a function of predicted risk. Additionally, by identifying ex ante which patients ought not to be tested, algorithmic predictions pave the way for targeted interventions to eliminate inefficient tests prospectively.

We next turn to the untested, where recent research highlights an important counter-point to over-testing: physicians may also under-test (Abaluck, Agha, Kabrhel, Raja, and Venkatesh, 2016). Because doctors weight certain variables differently than a structural model of risk would, some patients flagged as high risk by the model are left untested. We replicate this finding in our data: in the highest bin of predicted risk, where testing would appear to be very cost-effective, 42% of patients go untested. This fact raises the possibility of under-testing—but does not fully establish it. The key empirical problem is that we do not know what *would have happened* if we tested these patients. Contrast this with the case

of over-testing above: when the algorithm flags a patient as low-risk and yet the physician chooses to test them, we do not have to assume the physician was incorrect because she disagrees with the model. Instead, we look at actual test outcomes to judge the choice. We have no such recourse for the untested. When the algorithm flags a patient as high-risk, yet the physician chooses not to test, there is no test outcome to establish whether that choice was correct or not. We know only that the physician deviated from a statistical risk model. In fact, her deviation could be efficient. The physician has access to host of information unavailable to the model: how the patient looks, what they say, or in some cases even crucial data such as x-rays or electrocardiograms (ECGs).[3] While these high-dimensional data are not routinely included in predictive models, a physician could easily have seen something in the ECG that, comparing two otherwise similar patients, leads her to test one and not the other. For patients with ECG data available, we show this directly. Several ECG features predict both the physician's test decision and the yield of testing. And if we directly incorporate the high-dimensional ECG waveform into our risk predictions, using a deep learning model, it leads to measurable decreases in model-predicted risk for 97.5% of patients—and 100% of the highest-risk untested. This is just one instance of a more general principle: we cannot simply assume physicians select patients for testing based only on observables.[4]

To address this challenge, we look to new data. The reason we wish to treat blockages in the first place is because untreated blockages have consequences. Those consequences can be measured. If high-risk untested (and thereby untreated) patients actually did have a blockage, they should go on to experience adverse events. The clinical literature has distilled this insight into a composite measure of 'major adverse cardiac events,' typically formed in the 30 days after the patient's ER visit: diagnoses of heart attack or urgent invasive treatments for heart attack, which we confirm by cross-referencing them with concurrent laboratory biomarkers of heart damage; or cardiac arrest and death. Widely-used clinical decision rules set risk threshold values based on this metric, to establish which patients should have been tested: those who go on to have event rates above these values.[5] We find that in the highest-risk bin, untested patients go on to have an adverse cardiac event rate of 9.16%, high enough for guidelines to conclude they should have been tested. This is not just based on health records, which may reflect in part discretionary utilization of care: death accounted for a large fraction (3.82%) of these adverse events. As another point of comparison, if we take the riskiest 1834 untested patients—the size of the tested cohort—the

[3]Physician notes do not reliably capture all these data elements, and because they are often written days after visits, their content cannot be used in prediction algorithms.

[4]This selection bias pervades many machine learning applications, despite growing recent attention to the 'selective labels' problem (Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan, 2017). A related problem arises when measured outcomes are selectively changed by treatment, as opposed to testing which selectively reveals them (see Paxton, Niculescu-Mizil, and Saria (2013) for a thorough discussion).

[5]Such decision rules (e.g., TIMI, GRACE, HEART) are commonly implemented in ER testing protocols, as well as in eligibility guidelines for diagnostic technologies (e.g., CT-angiography) or treatments (e.g., statins). Importantly, we do not take a stance on whether these rules are physiologically optimal, only that they represent current physician understanding of who should be tested. If physicians are failing to test apparently high-risk patients because of private information, and applying the existing clinical knowledge, these high-risk patients should have rates below these thresholds, particularly given incentives to over-test.

total adverse event rate is 5.18% (of which death made up nearly half: 2.51%). In contrast, lower risk bins have much lower rates, well below the threshold for testing.

These adverse event data, though, do not rule out another source of physician private information: whether a patient is suitable for testing or treatment. Treatment for heart attack is invasive, and may be inappropriate for patients who are too frail or have other conditions that make procedures risky. Even diagnostic testing (i.e., catheterization) itself has risks, meaning physicians might hesitate to even order a test. As such, some truly high-risk patients will go untested, and some of these will go on to have an adverse event—but this need not be an error: the physician saw that the benefit of testing or treatment was limited. We have partially addressed this in our sample construction. Patients in whom the physician suspects heart attack, but who are not tested, will have this noted in their ER documentation, and we have excluded all such patients from calculations about our untested population above.[6] Still, we might be concerned about some residual private information about patient suitability, even after this exclusion.

To address this, we note that a patient suspected of a blockage, even if ineligible for testing or treatment, should not simply be sent home. At a minimum, they should have an ECG: a low-cost, non-invasive test performed on anyone in whom the physician has even a low suspicion of a blockage. Even for treatment-ineligible patients, the ECG will guide delivery of medications that non-invasively target coronary blockages (e.g., blood-thinners), and the intensity of cardiovascular monitoring (e.g., in the ICU). This implies that, if the adverse event rates we observe are driven largely by recognized treatment-ineligible patients, removing those with an ECG should address this problem: patients without an ECG are unlikely to be suspected of heart attack, let alone suspected strongly and additionally ruled out as too risky for testing or treatment.[7] We find that removing these patients does reduce the adverse event rate, suggesting some amount of private information. But at 5.66%, adverse event rates in the top bin remain above the clinical threshold mandating which patients should have been tested.[8] 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.

These two pieces of evidence rule out two potential rationalizations for why high-risk patients go untested: physician private information about risk; and physician private information about suitability for testing or treatment. But we have not presented any direct evidence that testing these patients would have yielded health benefits. This is important: we may have failed to address some other unsuspected source of unobservable difference in treatment benefit. Alternatively, our very definition of risk rests on the assumption that a positive test result followed by treatment indicates a good outcome. But if physicians over-

[6]We exclude any untested patient with a positive troponin lab, indicating known heart attack, and those in whom the physician notes diagnosis codes related to heart disease. Such documentation is medically useful, for future care, and economically prudent: ER visits are reimbursed proportional to medical complexity.

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

[8]We perform a similar analysis for patients lacking a troponin result, another low-cost test in the ER, and find rates of 6.80% in the riskiest bin.

treat, some of those treatments would have failed to improve health, meaning these 'high-risk' patients may benefit less than we think. Both of these concerns could be addressed if we could more directly measure the effect of additional testing on health outcomes.

To accomplish this, we rely on (somewhat) exogenous variation in who is tested. ERs operate in shifts, and different shifts test at different rates. So depending on the particular team working at the triage desk, for example, a high-risk patient coming in on a busy Monday evening will have anywhere from a 4.3% likelihood of being tested to 24.6%.[9] This natural variation simulates an experiment, in which a patient is more or less likely to be tested, for reasons unrelated to their actual risk. On average, across all patients, we see that increasing testing in this way has no statistically significant effect on health outcomes, matching what is often called 'flat of the curve' health care: more testing, little return. But as before, this average hides a great deal of heterogeneity by predicted risk. High-risk patients—and only high-risk patients—assigned to higher testing regimes do significantly better: 29.6% (4.7 p.p) fewer downstream heart attacks, arrhythmias, and death over the year after visits (with mortality accounting for a full 31% of this reduction). Cost-benefit calculations suggest that testing these high-risk patients is highly cost-effective, with a marginal cost-effectiveness of scaling up testing of $11,349. These results of course do not show that testing *all* high-risk patients would yield benefits: we cannot make inferences outside of the observed range of variation in testing rates across shifts. But they do establish the presence of a sizable set of high-risk patients whom physicians typically leave untested—and in so doing, leave large health gains on the table. As a whole, these results suggest physicians both over and under-test. The magnitudes of both appear to be large: under our preferred set of policy counter-factuals, addressing over-testing would cut 49.1% of tests, while addressing under-testing would add back in tests equal to 37.3% (of the current tests).

So far, we have documented the welfare costs of physician errors, but not answered an important question: why do physicians make these errors? To shed light on this, we build another predictive model. Rather than predicting patient risk, it instead predicts physician decision-making: for each patient, what is the probability that the physician will choose to test. [10] We first examine whether bounded rationality—limits in cognitive resources such as attention, memory or computation—could prevent physicians from approximating a full algorithmic risk model, which in our case is built up from 16,381 variables (Simon, 1955; Gabaix, 2014; Sims, 2003; Gabaix, 2017; Mullainathan, 2002; Bordalo, Gennaioli, and Shleifer, 2017). To this end, we build a set of risk models that vary in complexity (e.g. using more or less variables, or more or less iteration of statistical learning), and use these to predict who gets tested. We find that the risk model that best predicts whom the

[9]Patients' observable characteristics largely appear balanced, after accounting for variation in testing rates over time (e.g., by hour of day, day of week, etc.) and observable risk. In addition, we find no significant difference in realized yield across shifts conditional on predicted risk, arguing against large differences in unobservables (e.g., if higher testing rates were driven by higher unobserved risk, these shifts should have higher realized yield, but they do not).

[10]This approach builds on a long history of research comparing clinical vs. actuarial decision making to gain insight into physician decision making (Ægisdóttir et al., 2006; Dawes, Faust, and Meehl, 1989; Elstein, 1999; Redelmeier, Ferris, Tu, Hux, and Schull, 2001).

physician will test is much simpler than the one that best predicts who will have a coronary blockage. While the model is too simple, it is nonetheless an effective model of risk—and moreover, considering the variables that enter into it, physicians appear to weight them more or less correctly. To make an analogy with machine learning algorithms, it is as if the physician 'over-regularizes' (Camerer, 2018). These two findings provide some support for both boundedness and rationality. But this exercise also reveals particular variables that appear to play an out-sized role in the physician's testing decision, above and beyond objective risk. In particular, we find that a patient's symptoms, which are highly salient (as opposed to say their past history), are over-weighted relative to other inputs. For example, patients with chest pain are riskier than other patients—but they are also much more likely to be tested above and beyond their true risk. Following this logic, we investigate for systematic biases by asking what variables predict testing above and beyond true risk. In particular, we find the most salient aspects of risk - a patient's symptoms and demographics - are over-weighted. (Bordalo, Gennaioli, and Shleifer, 2012). In addition, we also find that symptoms which are representative of heart attack - i.e., common in heart attack patients relative to their base rate - are also over-weighted. While such symptoms are of course diagnostic, we find they are given excess weight above and beyond their diagnosticity (Kahneman and Tversky, 1972; Bordalo, Coffman, Gennaioli, and Shleifer, 2016). Together these results suggest both boundededness and systematic biases play a role.

Our results add to recent empirical work that argues for a richer model of physician decision making that goes beyond simple moral hazard. Chan, Gentzkow, and Yu (2019) highlight the importance of skill differences: some physicians are better than others in deciding whom to test. Chandra and Staiger (2020) highlight comparative advantage: some hospitals specialize and focus on certain tests and conditions. These papers, for different reasons, caution against a common policy prescription in health: encouraging high-testing providers to behave more like low-testing ones. This will work well if the problem is incentives—but can have perverse effects if other mechanisms are at play. In our case, physician error creates effects analogous to behavioral hazard in models of consumer choices: incentives to reduce testing may be too crude a tool, simultaneously reducing both wasteful and highly useful tests (Baicker, Mullainathan, and Schwartzstein, 2015). We show this directly, in both electronic records data from the hospital we study, and in national Medicare claims. In both cases, low-testing physicians (or hospitals) test less *across the entire risk distribution*. They are in effect throwing the baby out with the bathwater, sacrificing both high-value and low-value care alike.

2 Data and Approach

2.1 Medical Context: Testing for Heart Attacks

Heart attack is a colloquial term for acute coronary syndrome (ACS): reduction of blood flow to the heart, due to a new blockage in the coronary arteries supplying it. This leads to damage or death of a patch of heart muscle, which has both immediate consequences,

e.g., sudden death from arrhythmia, and longer-term sequelae, e.g., congestive heart failure, chronic pain. Heart attack is treated with 'revascularization' procedures to open up blocked coronary arteries, typically using a flexible metal tube called a stent to relieve the blockage (or less commonly, with open-heart surgery). These interventions can be life-saving: decades of large and robust randomized trials in the emergency setting have shown a dramatic effect on both mortality and morbidity (reviewed in Amsterdam et al. (2014)).

But in order to treat a blockage, one must first diagnose it. For the 139 million patients every year coming into US emergency departments (EDs), this task falls to the emergency physician. It is easier said than done. Patients with life-threatening blockages can have a range of subtle symptoms, e.g., a subtle squeezing sensation in the chest, shortness of breath, or just nausea. These symptoms are common in the population seeking emergency care, and most often result from benign problems in other organ systems: acid reflux, viral infections, or a pinched nerve in the back (Swap CJ and Nagurney JT, 2005). Not every patient can consult with every potentially relevant specialist ex ante—a cardiologist, gastroenterologist, infectologist, orthopedist, etc—to determine the source of the problem. So the emergency physician must make the first set of testing and triage decisions, with the goal of arriving at a diagnosis, or at least, excluding the most concerning diagnoses. To give a sense of the pace of decision making, physicians in the ED typically see 2-3 patients per hour over shifts that last 8-12 hours. As part of her initial decision making, the physician can order simple tests in the ED (the electrocardiogram, and laboratory tests like troponin, described in more detail below). These data can help her triage patients, in terms of their likelihood of blockage and the urgency with which they must be treated. But no test done in the ED can actually diagnose a blockage. For that, she needs to either hold the patient for further testing, or consult a cardiologist for advice.

The definitive test for acute coronary blockages is cardiac catheterization, an invasive procedure carried out in a dedicated 'catheterization laboratory' (a facility separate from the ER). A cardiologist inserts an instrument directly into the coronary arteries, squirts in radio-opaque dye, and visualizes the presence and location of the blockage via x-ray. Critically, this procedure also allows for the delivery of the treatment: a 'stent' or flexible metal tube that restores flow through the artery.[11] An alternative testing pathway adds a step before catheterization, called 'stress testing.' The intuition for these tests will be familiar to economists: heart muscle has some demand for blood, which is supplied via the coronary arteries. If supply is limited by a blockage, it may not be apparent at rest, when demand is low. So to identify limitations in supply, the stress test creates a shock to demand for blood, e.g., by asking the patient to exert herself on a treadmill or by administering a drug to increase heart activity. By monitoring the heart's response during this procedure, supply-side limitations become obvious, suggesting a blockage. The advantage of the stress test is that, if negative, an invasive procedure has been avoided; the disadvantage is that, if positive, the patient still needs catheterization – and precious time has been wasted.

[11] A less common treatment is to bypass the blockage, by surgically adding a new blood vessel – specifically, transposing it from the chest wall or the ankle to the heart – via open heart surgery. This is only offered to those with multiple blockages revealed on catheterization, so in practice catheterization is still a prerequisite for treatment.

Whatever pathway the physician chooses, the proliferation of testing has been part of the dramatic reductions in rates of missed heart attack in the ER, which in the 1980s and 1990s were substantial: anywhere from 2-11% of heart attacks (Pope et al., 2000; Schor S et al., 1976; Lee et al., 1987).

Of course, these tests also have costs. Financial costs are both direct – thousands of dollars for stress tests and tens of thousands for catheterization – and indirect, arising from the need for overnight observation and monitoring before testing can proceed. There are also health costs. Of all imaging tests, stress tests carry the single highest dose of ionizing radiation, which is thought to substantially raise long-term cancer risks (Mettler, Huda, Yoshizumi, and Mahesh, 2008; Brenner et al., 2003).[12] More directly, the definitive test, cardiac catheterization, is an invasive procedure: in addition to a large dose of ionizing radiation, it involves injection of intravenous contrast material that can cause kidney failure, and a risk of arterial damage, and stroke (Betsou, Efstathopoulos, Katritsis, Faulkner, and Panayiotakis, 1998; Rich and Crecelius, 1990; Katzenschlager et al., 1995; Hamon, Baron, Viader, and Hamon, 2008).

2.2 A Simple Framework

To structure our analysis, we rely on a simple framework that captures the key aspects of testing for heart attack. A patient appearing in the emergency room might or might not have a blockage denoted by $B = \{0, 1\}$. The physician can use a diagnostic test denoted by $T = \{0, 1\}$ to detect the blockage, at cost c_T. If a blockage is present, the test detects it with probability q. If there is no blockage, it will come back negative.[13] A physician can also deploy a treatment (most commonly, stenting), $S = \{0, 1\}$, that will address the blockage, at cost c_S.

Not all patients pay the same health costs for testing and treatment. In particular, we assume that some patients, because of frailty or other medical conditions (e.g., bleeding risk), cannot handle invasive procedures like catheterization. As a result, these patients pay a higher health cost of testing, that reflects both the direct effects of catheterization, but also those resulting from stress testing (in expectation, because positive stress tests lead to catheterization at some rate). They also get less benefit from treatment if a blockage is present. We will identify these patients with variable $K = \{0, 1\}$ that flags a contraindication to invasive procedures. We assume that K captures heterogeneity in both the health costs of treatment and of testing.[14] Our goal here is to capture the *known* heterogeneity in the

[12] Exercise on a treadmill in the setting of heart attack, as required for traditional stress testing, is thought to pose a "small but definite" risk of cardiac arrest (Cobb and Weaver, 1986). This may simply reflect confounding, since patients are tested because of suspicion for heart attack, which itself carries a risk of cardiac arrest.

[13] We account for the effect of false positives below, in our discussion of treatment effects. To be clear, we are not assuming away false positives of testing, simply incorporating them via their influence on treatment effects, because they have the same net consequence: false positives (people without heart attack who are treated as if they had one) will reduce the expected health benefits of treatment. We provide sensitivity analyses with various treatment effects in our cost-effectiveness analysis.

[14] This assumption is reasonable because testing and stenting are the same procedure, and the health costs

costs of testing and treatment. While there may be other forms of heterogeneous effects due to physiology that are yet to be discovered, the contraindications we model are the ones understood by clinicians and codified in clinical guidelines. Because our goal here is to study physician choice, as benchmarked against existing knowledge, we will focus on this form of heterogeneity. We discuss the details of ascertaining this variable in the data below.

Each patient has health W as well as some characteristics at the time of their visit denoted by X, Z. Physicians can see both X and Z and use them to form their judgments, whereas only X is measured in the data. We assume that expected health given these characteristics equals

$$E[W|X, Z, S] = w(X, Z) - b(X, Z)(\eta - \tau S)$$

where w and b are deterministic functions of X, Z that identifies baseline health W in the absence of blockage, and blockage B, respectively. A blockage negatively impacts health by η, and the key contribution of the physician is to allocate the treatment S to offset this loss by τ (with $\tau \leq \eta$).

2.2.1 Optimal Testing Strategy

In allocating testing, society's objective is to choose testing and treatment (T and S) to maximize expected health net of costs, which we write as:

$$E[W|X, Z, S] - c_T T - c_S S - (\delta_c T + \delta_\tau S)K.$$

The first two costs, $c_T T$ and c_S, are financial, and paid by all tested patients. The last term captures the additional health costs paid by patients with contraindications, $K = 1$, for both testing δ_c and treatment δ_τ. (Here we assume that K is a subset of X, and that physicians ascertain it accurately; we return to the details of measurement, and describe analyses to check the robustness of these assumptions, below.)

The additional costs paid by contraindicated patients have an important implication for the optimal testing strategy, which can be seen if we work backwards from the treatment decision. For illustrative purposes, we begin with the case of a patient in whom a a test reveals a blockage. For such a patient, the net expected benefit of treatment is equal to $\tau - \delta_\tau K - c_S$. We interpret current medical guidelines and practice as consistent with three assumptions. First, we assume that $\tau - \delta_\tau < c_S$, so contraindicated patients are never treated (even without incorporating any additional costs of testing these patients). Second, we assume that $\tau - c_S \geq c_T$, so that blockages in eligible (non-contraindicated) patients will always be treated. Third we assume that for eligible patients, $Pr(B = 1|X, Z)(\tau - c_S) < c_T$ so that in the absence of a definitive test showing blockage, the physician will not treat (i.e., no presumptive treatment without confirmatory testing; in any case, treatment mechanically requires catheterization to deliver the stent to the location of the blockage).

laid out in canonical medical texts are largely driven by procedural complications (Baim and Simon, 2006). It is also supported empirically in studies showing that rates of complication in diagnostic vs. therapeutic catheterization are correlated, but treatment complications rates are higher, (Chandrasekar et al., 2001; Wyman et al., 1988).

This in turn implies an optimal testing rule, where again the goal is to simply maximize expected health minus costs. We define this expected benefit of testing a patient with characteristics X, Z as:

$$V(X,Z) = Pr(B=1|X,Z) \times q(\tau - c_S - \delta_\tau K) - (c_T + \delta_c K)$$

For those with $K=1$, this is strictly negative so they will not be tested: even with a positive test they would not be treated, so why test? For those with $K=0$, the value of testing is positive as long as the risk of a blockage is high enough. Specifically if $Pr(B=1|X,Z) > \frac{c_T}{q(\tau - c_S)}$ it is optimal to test. This results in a testing rule:

$$T=1 \text{ iff } K=0 \text{ and } Pr(B=1|X,Z) > \frac{c_T}{q(\tau - c_S)}$$

Patients with high enough risk of a blockage are tested if they have no contraindications; patients with contraindications are never tested however risky.

2.2.2 Physician Testing Strategy

Physicians of course need not follow this rule, for two reasons: they may have different objectives, or inexact beliefs about risk. First, we assume that physicians may have a private additional benefit ν to testing.[15] This term captures moral hazard and incentives to over-test. Second, we assume that their judgment of a probability of a blockage equals $h(X,Z)$, which can differ from true probabilities ($Pr(B=1|X,Z)$), but need not. The result is that physicians do not value testing according to $V(X,Z)$ but instead according to $V_h(X,Z)$:

$$V_h(X,Z) = h(X,Z) \times q(\tau - \delta_\tau K - c_S) - c_T + \nu$$

Calculating the condition that physicians test if $V_h > 0$ yields the following testing rule:

$$T=1 \text{ iff } K=0 \text{ and } h(X,Z) > \frac{c_T - \nu}{q(\tau - c_S)}$$

Relative to the socially efficient rule, there are several distortions here. First, if $\nu > 0$, there will be too much testing, as in a moral hazard model of physician behavior, where physicians rank patients correctly by risk but set too low a threshold. Second, since $h(X,Z)$ need not equal $Pr(B=1|X,Z)$, testing decisions can fail to match patient risk, throughout the risk distribution. In a 'mis-prediction' model, there can be both too many and too few tests; or more precisely, the problem is not in the number of tests but in who gets tested.

[15] We have abstracted here from moral hazard and over-*treatment*, i.e, conditional on test results, for two reasons. First, our goal is to identify patients who benefit most from testing, subject to the constraints of current medical knowledge about treatment–we have nothing to say about which patients benefit from treatment. To the extent there is over-treatment in practice, these effects are already implicitly accounted for in our treatment estimates from randomized trials, which are 'diluted' by patients who are treated but do not benefit. Second, we believe the extent of moral hazard here is minimal: unlike the decision to test a patient, the decision to place a stent conditional on the results of catheterization is largely driven by objective data, on the level of blood flow through the coronary artery in question, and by clinical guidelines. This is not to say there is no moral hazard, but there is certainly less of a gray area than in testing decisions.

2.2.3 The Role of Algorithmic Predictions

Against this backdrop, we consider an estimator $m(X)$ of patient risk $Pr(B = 1|X)$. Notice this algorithm has less information than the physician as it only uses X and not Z (again, recall K is part of X, a point we return to later). We assume the algorithm is consistent and $m(X) \to Pr(B = 1|X)$; and that the algorithm is applied in a finite sample so that $m(X) \approx Pr(B = 1|X)$. These assumptions motivate the empirical approach to quantifying over- and under-testing, which we describe next: to measure outcomes related to realized blockage B, for patients above or below some threshold of $m(X)$. But we emphasize that neither assumption is necessary: our tests are properly sized without it, and the accuracy of the algorithm alone determines the power of our tests. An algorithm that fails to extract much signal for predicting B (or mis-predicts in general) will fail to detect errors in testing even when they exist; but the errors we do detect will in fact be errors.

Three points are worth making about the way we use algorithmic predictions. First, the algorithmic predictions themselves are not used as benchmarks. We are not assuming a physician is in error if their testing rule deviates from the algorithm. Rather, the algorithm is only used to identify candidate patient groups where there *might* be over- or under-testing. We then look to the data, to determine empirically whether the algorithm has provided useful information. Second, our definition of over- and under-testing is posed in terms of welfare losses, rather than any specific kind of physician error (e.g., bias, excess variance, etc). In effect, we are asking about the welfare effects of two counterfactuals: can we identify ex ante patient groups who are currently tested, but welfare would rise if they were untested; and can we identify ex ante patient groups who are currently untested, but welfare would rise if they were tested. In other words, we use the algorithmic predictions to test whether physicians are leaving signal on the table, and how much health gains they forgo as a result. Third, because we study physician decisions by contrasting them to algorithmic predictions on a test by test basis, we effectively allow for physicians to vary both in their thresholds for testing and in their ability. The physician has a risk-predictor $h(X, Z)$, and the nature of that risk-predictor determines their ability. This model is therefore quite general in allowing for skill differences and in different thresholds. For the bulk of this paper, though, we do not explicitly study these differences. Instead, conditional on whatever physician's skills may be, we ask whether an algorithmically derived predictor could meaningfully improve welfare. Because we do not need to make assumptions about individual physicians to construct our estimates of over- or under-testing, our results reflect average practices and skills by physicians in our sample; in principle they could be used to study decisions at the individual level, although this is not the focus of our study.

2.2.4 Over-testing

We first examine over-testing: whether physicians test patients more than is socially optimal. Our strong prior is yes, because over-testing as an empirical fact is well-documented in health economics. Incentives are the most common explanation: insurers pay by the test,

and malpractice lawsuits push to excess caution.[16] Studies of testing in particular, across a variety of different medical settings, have substantiated this model with a consistent empirical fact: low average yield. Testing for heart attack is a frequently cited example, because the cost of testing is high and yields are particularly low – often 1-2% (Foy AJ et al., 2015; Hermann et al., 2013; Rozanski et al., 2013; Stripe, Rechenmacher, Jurewitz, Lee, and Schaefer, 2013; Rokos et al., 2010). In light of this, health policy commentators have recently begun to advocate dramatically scaling back testing, particularly non-invasive stress tests, or in some cases eliminating it altogether (Prasad, Cheung, and Cifu, 2012; Redberg, 2015). More broadly, high-profile efforts to reduce low-value care focus heavily on over-testing: for example, of the American Board of Internal Medicine's "Choosing Wisely" list of low-value care items, 62% were aimed at reducing over-use of diagnostic tests (Morden, Colla, Sequist, and Rosenthal, 2014).

A key empirical observation that makes the study of over-testing possible in any domain, is an observation about counterfactuals. Since we know how tests turned out – and know that negative tests do not lead to health benefits – we can consider what would happen under a counterfactual policy: do not test some pre-defined group. We know the yield for this pre-defined group, and hence the benefits from testing we would forgo, as well as the cost of tests in this group.

Building on this, most studies measure over-testing by looking at yield of tests as a whole (Foy AJ et al., 2015; Hermann et al., 2013; Rozanski et al., 2013). Specifically, they take the entire sample of the tested, calculate the net benefits of each test and look at the average net benefit:

$$E[V(X,Z)|T=1] = Pr(B=1|T=1) \times q(\tau - \delta_\tau K - c_S) - (c_T + \delta_c K)$$

This incorporates known estimates of τ, the health benefits of stenting, and accounts for any (mis)treatment or testing of those with contraindications $(\delta_\tau + \delta_c)K$, as well as the financial costs of testing and treatment c_S and c_T. To put these into dollar units then requires only an assumption of the dollar value of a life year saved by τ .

Such estimates, by looking at the overall average yield of tests are weak in two senses. First, they are implicitly considering a coarse counterfactual. If we found $E[V(X,Z)|T = 1] < 0$, we would implicitly be arguing that testing no one yields higher welfare than the current testing regime. However true it might be, this is coarse and unrealistic as a counterfactual. Second, they may fail to detect over-testing even if present. Suppose that $E[V(X,Z)|T = 1] > 0$. It can still be that there is a set of identifiable patients $\mathcal{S}$ such that $E[V(X,Z)|(X,Z) \in \mathcal{S}] < 0$. Just because the average test is valuable does not mean that there are not large swathes of tests with negative value. Of course, this is related to the counterfactual point. Ideally we would like to identify which particular tests are wasted, and consider counterfactuals where those are cut. This more comprehensive approach would involve fully characterizing the distribution of $V(X,Z)$ and identifying the negative value parts of that distribution. Given that only X is observed, a realistic goal would be to map out

[16] See Greenberg and Green (2014) and O'Sullivan et al. (2018) for a review, and Acemoglu and Finkelstein (2008) and Baker (2001) for empirical evidence in economics.

$E[V(X,Z)|X=x,T=1]$. This is simply a generalization of the approach taken by studies in the literature that quantify the value of testing in particular cells of X ex ante researchers suspect to be low-value, e.g., patients under age 40 (e.g., Ely et al. (2013)). Unfortunately, given the huge number of such cells, directly calculating $E[V(X,Z)|X=x,T=1]$ for every X cell is infeasible.

A risk prediction algorithm $m(X)$ makes this exercise more tractable. Instead of looking at all X cells, we can calculate net value of testing for patients defined by some bin (or value) of predicted risk R. Specifically we define for each $m(X)$ bin in some range $r = [\underline{r}, \bar{r}]$ the value of testing to be:

$$V_m(r,t) = E[V(X,Z)|m(X) \in r, T=t]$$

The added $T=t$ condition is important here. Note that both the value of testing and the testing decision vary within an X cell: they both can vary by Z. As a result, the average value of testing in a given X cell can differ between the tested and untested. Because of this $V_m(r,T=1) \neq V_m(r,T=0)$. In a given predicted risk cell, the untested and tested can have different returns to testing.

This caveat will have large implications for our analyses in the untested below, but it does not prevent us from characterizing the value of testing in the tested: the data elements we need are all present in our dataset. Concretely, after mapping out the distribution of $V_m(r,T=1)$ we would look for ranges of $m(X)$ where the function is negative, but physicians nonetheless test at meaningful rates. We thus look for over-testing in these cells, by calculating

$$V_m(r,T=1) < 0 \ , \ Pr(T=1|m(X) \in r) > 0$$

If we found such a set then we have found a counterfactual that would increase welfare: cutting all tests with $\underline{r} \leq m(X)$. In other words $m(X)$ lets us identify subsets of the patient population that are potentially low-value. This is a more precise counterfactual than cutting all tests, which is the only tractable option in the absence of $m(X)$, when we are restricted to looking at aggregate yields alone.

There are three points worth making about this approach. First, note that it is robust to the physician's information advantage. Certainly, the algorithm's risk prediction relies on X whereas the physician has access to both X and Z. But despite this, we can make an inference about whether the physician is over-testing. Suppose we find a value r where, for patients with $m(X) \in r$, physicians sometimes test, $Pr(T=1|m(X) \in r) > 0$; but the net value of testing is negative, $V_m(r,T=1) < 0$. Since $V_m(r,T=t) = E[V(X,Z)|m(X) \in r, T=t]$ it follows automatically that there must be some X,Z where $V(X,Z,T=1) < 0$: if a whole a cell shows negative value of testing, then at least some patients in that cell must have negative value of testing. And while there might be patients in that cell who have $V(X,Z,T=1) > 0$ and should have been tested, this does not contradict the claim that, on average, we would be better off not testing this cell. Second, note that our ability to detect over-use is only as powerful as the algorithm $m(X)$. A very bad algorithm could fail to find over-testing even if there are cells in which over-testing occurs. The test is only as good as the estimator. Absent access to the true $V(X,Z)$ function, however, this will always be a

weakness. Finally, and conversely, note that if we find such cells using any $m(X)$ function we would have evidence of over-testing—however that $m(X)$ was defined. Putting these points together, the power of our approach (whether it finds over-testing) depends on the quality of the risk estimator, but the consistency does not: if it isolates a cell with negative expected value, it has identified over-testing. Finally, as we've emphasized in this whole discussion, the only role of the algorithm $m(X)$ is to identify subsets of patients who are *potentially* low-yield. But we verify these estimates of whether this subset patients should have been tested based on *realized* yield—not on the value of $m(X)$. The algorithm only identifies plausible subsets in which we can look for empirical evidence of over-use.

2.2.5 Under-testing

We next turn to the question of whether some patients who were *not* tested should have been. In particular we are looking for an X, Z cell such that $V(X, Z) > 0$ but with $Pr(T = 0|X, Z) > 0$. As before, because we rely on an algorithm with access to only X, we are looking for cells where

$$V_m(r, T = 0) > 0 \text{ and } Pr(T = 0|m(X) \in r) > 0$$

Such counterfactuals, though, are far more difficult than those we have dealt with so far: finding patients who *were* tested and should not have been. For tested patients, we have the test results. But for untested patients, we do not know what the test result would have been. Because of this asymmetry, we cannot calculate $Pr(B = 1|T = 0, m(X) \in r)$ as we could in the tested, and hence cannot calculate $V_m(r, T = 0)$. A common solution to this problem is to note that if the physician testing rule only depends on X and not Z, then it follows that $V_m(r, T = 0) = V_m(r, T = 1)$. Of course, such an assumption—that the algorithm and the physician have the same information set—is both unrealistic and builds in an unfair advantage for the algorithm: if we assume the physician and the algorithm have access to the same variables, it is hardly surprising that an algorithm built on dataset with many more observations will form a better predictor. A more robust examination of under-testing must allow for the possibility that $V_m(r, T = 1) \neq V_m(r, T = 0)$, and assess the extent of the physician's information advantage empirically, rather than assuming it away.

To make headway, we exploit a fact about the physiology of heart attack. So far we have relied on a positive test result, with some probability q, to reveal blockages: hence the test yield, in terms of treatments S, is a proxy for B. But untreated blockages also have consequences that manifest over time – indeed, this is why we test and treat blockages – and these can also serve as proxies for B. Knowledge of the specific health consequences have been mapped out in a long tradition of clinical research. We draw on this to form A, a binary variable that captures whether a patient suffers these adverse events after the original visit.[17] Of course these adverse events can happen to patients who do not have a blockage.

[17]For the sake of simplicity, we postpone discussion of the specific outcomes, and the time window over which these adverse events occur, to the empirical results below. We choose both the basket and outcomes and the windows based on clinical knowledge, and show effects for different outcomes and windows.

Similarly a patient who has been treated will be less likely to have an adverse event. We thus assume that:

$$Pr(A = 1) = \gamma + Pr(B = 1) \times (1 - \phi S)$$

Here γ is the base rate of an adverse event absent a blockage and ϕ indicates the effect of stenting (the treatment) on reducing the chances of an adverse event.

Previously we derived a test as being valuable if $Pr(B = 1|X, Z) > \frac{c_T}{q(\tau - c_S)}$, using a threshold that captures the cost-effectiveness of treatments in tested patients. The clinical literature provides a similar threshold in untested patients, that captures the maximum acceptable 'miss rate' in terms of adverse events. Such bounds are used in a variety of medical contexts, described in detail below, from clinical guidelines for testing to allocation of new diagnostic technologies for heart attack. Specifically, if a set of untested (and thuse untreated and undiagnosed) patients go on to have realized adverse events A at higher than rate θ (typically 2% in the 30 days after visits) ex post, they should have been tested. As a result we know that a test has high value if

$$Pr(A = 1|X, Z) > \theta$$

The key to this threshold is that it reflects existing medical knowledge: physicians would (and should) test a patient if they knew the adverse event risk was above θ.[18]

We can use this to identify under-testing. Specifically, we first look for predicted risk cells where testing is predicted to have high value, yet patients are not being tested: $V_m(r, T = 1) > 0$ and $Pr(T = 0|m(X) \in r) > 0$. Of course in these cells it is possible that $V_m(r, T = 0) < 0$ so that testing the untested here does not make sense. So then, using the above logic, we test for under-testing in these cells by calculating

$$Pr(A = 1|m(X) \in r, T = 0, K = 0) > \theta \ , \ Pr(T = 0|m(X) \in r) > 0$$

Are adverse event rates in untested patients above the threshold that would suggest they should have been tested? Note that to call this under-testing, we require these untested patients to also have no contraindications to testing or treatment. If patients are ineligible, adverse events would not be evidence of error – they would be inevitable: yes, we would have found a blockage if we tested, but the patient would not have benefited from treatment (or been harmed by invasive testing). This is why we have stipulated $K = 0$. In practice, when we look at the untested, we will exclude these patients in our baseline specifications, in several specific ways detailed in Section 3.2, and provide careful robustness tests to ensure we have excluded the $K = 1$ population.

2.2.6 Physicians' Testing Margin

So far, we have used algorithmic risk predictions to quantify over- and under-testing, and thus evaluate the quality of physicians' decision making processes. But these predictions also

[18]Note that we do not assume this threshold is socially optimal; we do assume that physicians believe it to be optimal, on the basis of published guidelines, for the purposes of defining under-testing.

let us go deeper into the behavioral underpinnings of these decisions, by providing empirical tests of predictions made by different models of physician behavior.

Recall the physician's objective function $V_h(X, Z)$, which incorporates both the physician's predicted risk $h(X, Z)$ and any private additional benefit ν she receives for testing. Under a classical moral hazard model of physician behavior, incentives influence ν (such that $\nu > 0$), but not the physician's risk predictions $h(X, Z)$. Let us assume physicians correctly use observable information: $h(X, Z) = Pr(B = 1|X, Z)$. Assume the algorithm also correctly uses risk information $m(X) = Pr(B = 1|X)$. A model in which patients are correctly ranked, but the threshold is set too low, makes three empirical predictions. First, we should find a range of patients with negative return based on $m(X)$ alone, but where physicians nonetheless test with some nonzero probability, because the negative return is outweighed by ν. Second, below that range of $m(X)$, physicians will fail to consistently test, and these decisions not to test are efficient. Third, above that range of $m(X)$, physicians will consistently test, and these decisions to test are also efficient.

By contrast, consider a different model, where $h(X, Z) \neq Pr(B = 1|X, Z)$ (and is agnostic regarding the distribution of ν). This model makes a similar empirical prediction to moral hazard concerning patients with negative return based on $m(X)$: we should find that some of them will be tested by physicians. While the empirical prediction is the same, the mechanism is very different: rather a change in the testing threshold due to ν, here physicians mis-predict risk in low-risk patients, whom they perceive to be riskier than they are. Critically, such a model also makes a second empirical prediction: we should find a large number of patients with positive return based on $m(X)$ who go untested, because physicians perceive them to be less risky than they are. This is where a more behavioral model differs sharply from classical models of physician decision making. Under moral hazard, these patients should be tested at very high rates: they are highly profitable, because they are the most likely to generate the complex procedures and intensive care needs that are major contributors to hospitals' bottom lines (Abelson and Creswell, 2012); they are certainly more profitable than a negative test. So a rational physician would never prefer to test a low-risk patient ahead of one of these high-risk patients. The behavioral model results in a particularly inefficient outcome: low-value patients are tested more than they should be, as in moral hazard, but high-value patients are also left untested.

Our strategy for distinguishing between these conflicting predictions begins with a situation where we observe variation in testing as a function of some plausibly exogenous factor G. In the simplest case, consider two testing regimes: $G = 0$ when fewer patients are tested and $G = 1$ when more are tested, for reasons unrelated to X, Z. By identifying the marginal patients, who are tested when $G = 1$ but not when $G = 0$, we can directly inspect which cell of $m(X)$ physicians are drawing patients from. The key empirical difference will be found in the high-risk cells: if these contain *no* marginal patients, because all of them are already being tested in the low-testing regime, this would support a moral hazard model. More generally, marginal patients should have strictly lower risk than those tested when $G = 0$. If, on the other hand, we find increases in testing rate for high-risk patients – or haphazardly, throughout the risk distribution – we would conclude that the mechanism is more likely to

be related to physicians mis-ranking patients by risk.

Note that, while it is not the main emphasis of our paper, this framework is generalizable and can be used to test for differences across physicians. Since the model above is specified at the level of patient-tests, we can pool across tests to estimate parameters at the individual physician level, analogous to structurally estimated parameters in the literature, for example Abaluck, Agha, Kabrhel, Raja, and Venkatesh (2016). We can also compare these parameters to physician-specific skill levels as in Chan and Gruber (2020).

2.3 Data

Our primary data source for the main analysis is the data warehouse containing electronic health records (EHRs) from a large urban hospital. The dataset contains a wealth of information that gives us a window into physician decision making, and patient risk information and outcomes. We distinguish between five categories of observed data.

First, we observe demographics, including age, sex, race, and zip code. Crucially, we also obtain data on mortality, via linkage to Social Security mortality records, which capture death even if it happens outside of the hospital. Second, we also have a large set of structured diagnosis and procedure codes (ICD-9, HCPCS). These are present in both EHR as well as the national Medicare claims data we describe below, and record encounters between a patient and the health care system at points where payment is exchanged (e.g., a visit to a cardiologist to check cholesterol). They are mostly sourced from financial transactions between the hospital and the insurer, with some medical content – the diagnosis and procedure codes – and parallel datasets derived from insurance claims (e.g., Medicare data).

In addition to demographics and structured codes, electronic health record data additionally contain a wealth of information recorded in the course of routine clinical care. As a result, they contain three additional categories of data, that give us more granular insights into both patient risk and the physician's thought process and actions. The largest of these categories is quantitative biomarkers: laboratory studies (e.g., prior cholesterol results, biomarkers of heart injury) and vital signs (e.g., blood pressure, heart rate). Next, we see all medications in the patient's record, both those prescribed in the outpatient setting and those administered in health facilities. Finally, we see highly granular information from the ED visits we study: in particular, the patient's symptoms ('chief complaint'), as recorded by the triage nurse at the ED front desk, all the orders written by the physician during the visit, and the note they write to summarize that encounter.

2.3.1 Sample

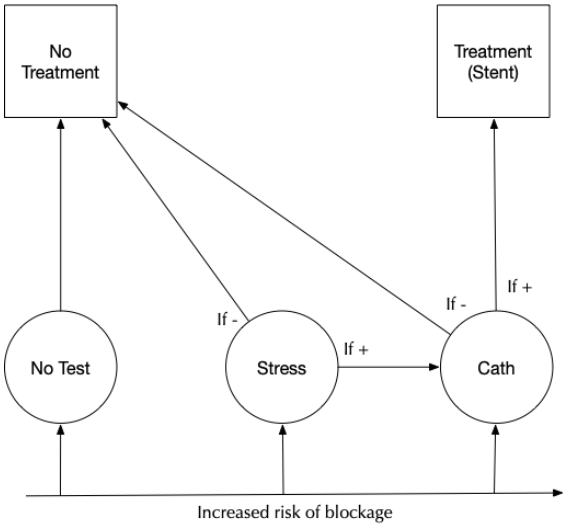
We use these data to form our sample: a consecutive series of 326,126 ED visits (indexed by i), by 150,862 patients (indexed by j), over a five and a half year period from January 2010 through May 2015. Each ED visit becomes an observation in our dataset. We exclude visits in which the patient died in the ED and thus before they could be tested (0.07%); visits preceded by recent known heart attack or its treatment (e.g., catheterization, stenting in the 30 days prior to ED visits: 0.2%) for whom testing may represent follow up of a

known problem, rather than diagnosis of a new one; and patients whose general poor health might limit the benefit of testing or treatment: patients 80 years of age or older (7.4%), those with poor-prognosis conditions diagnosed in the year prior (16.6%, e.g., with known metastatic cancer, dementia, on hospice or nursing home care, etc.).[19] Summary statistics on demographics and concurrent medical illnesses for the final sample of 246,874 ED visits, by 130,059 patients, are shown in Table 1. Of all visits, 2.9% were tested, 33.4% with immediate catheterization and 66.6% with stress tests (of which 13.0% subsequently had catheterization). On average, tested patients were older and had more risk factors for heart attack (prior heart disease, diabetes, high blood pressure, cholesterol) than untested patients. They were also more likely to be white and male.

2.3.2 Testing

By 'testing for heart attack' we mean testing for an acute blockage obstructing blood flow through one or more coronary arteries. Two types of tests can be used to diagnose these blockages; practically, we group these together in our definition of testing, but in this section we delve into some of the nuances of testing relevant to our subsequent analyses. First and most straightforward is catheterization, an invasive procedure that is both the definitive test, and also the means by which treatment in the form of stenting is delivered. Catheterization can either be done as the initial test, or the physician can choose to start with a stress test: a set of non-invasive procedures like putting the patient on a treadmill with electrocardiographic monitoring to look for signs of blockage, or radiological studies that visually quantify potential blockages. Importantly, while stress tests can suggest blockages, they cannot confirm or treat them. So if the physician chooses to first perform a stress test, it functions only as a first, lower-cost step to screen out negatives. If positive, the patient must proceed to catheterization for confirmation and definitive treatment.

The costs and benefits of these two categories of tests are different. Because it is an invasive procedure, the cost of catheterization is high. Financial costs total nearly $30,000, and there is a small but measurable set of procedural risks (most catastrophically, stroke). The benefit, of course, comes in the form of treatment for heart attack, if the catheterization identifies a blockage. Stress tests, because they are non-invasive, have lower costs: financial costs are around $4,000, and the health costs are minimal. Of course, if they come back positive, the patient will go on to pay the costs of catheterization, as well as the cost of the stress test. The benefit of stress tests is thus also lower: because they



[19] See Shanmugam, Harper, Meredith, Malaiapan, and Psaltis (2015) and Obermeyer, Cohn, Wilson, Jena, and Cutler (2017) for rationale and additional details on exclusions.

are not perfect, (i.e., the probability they will be
positive is less than 1) the joint probability of both a stress test and a catheterization will be positive is less than the probability that catheterization alone would be positive. So while they lead to the same benefit when both tests are positive, via delivery of treatment for heart attack, they are less likely to lead to benefit on average. As a result of this calculus, summarized in the diagram below, it is intuitive and efficient for physicians to begin with stress testing for lower-risk patients: their expected benefit—the likelihood of testing positive, going on to catheterization, and receiving beneficial treatment—is less than the cost of catheterization, but greater than the cost of the stress test. Higher-risk patients, on the other hand, should go straight to catheterization without incurring the additional cost and likelihood of false negative from stress testing.

Our analysis accounts for these nuances in two ways. First, because physicians use both types of test to answer the same question—does their patient have an acute coronary blockage—we group them together in our definition of testing T_{ij}. (By contrast, in other settings, largely outpatient clinics, these tests may be done for a variety of reasons: to characterize a baseline for future reference, to plan an elective surgery, etc.) We set $T_{ij} = 1$ if patient i has procedure codes for either stress testing or catheterization in the 10-day window (inclusive) after visit j. This window is based on guidelines for testing, which range from, e.g., 72 hours in Amsterdam et al. (2014) to 1-2 weeks in Brown et al. (2018). Practically, most tests (81%) are done either during the ED visit or in the 72 hours after. Importantly, to we allow the financial costs of testing to vary in our cost-benefit analysis, and thus keep careful track of which type of test is done on which patient. Overall, we identified 7,320 tested visits, and 242,343 untested visits (which are described in more detail below). Of the tested, 4,876 had stress tests, and 3,080 had cardiac catheterization; 636 had both a stress test and subsequent catheterization, indicating a positive stress test.

Second, the existence of different testing strategies means there is more than one counterfactual we might wish to consider in our analyses of testing decisions. For our main analyses, we consider a specific counterfactual: eliminating (or adding) the specific kind of test the physician chose to do in our dataset, in a patient with a given level of predicted risk. Other counterfactuals are also of interest, for example, eliminating all stress tests, or all catheterizations. We might also be interested in counterfactuals where we replace all catheterizations with stress tests, and vice-versa. We explore some of these in Appendix 4, where we first calculate the value of testing with catheterization or stress test alone, to simulate a scenario where one type of test is eliminated (added). We cannot fully simulate substitutions between testing types: for example, we do not know what stress tests would have shown in catheterized patients. But we can calculate a conservative bound on substitution of catheterization for stress tests, where we drop only negative stress tests (those that do not progress to catheterization). This simulates a near-perfect strategy of replacing current stress tests with catheterization.

2.3.3 Diagnosis of Blockage in the Tested

Among tested patients, if catheterization—whether done directly, or following a stress test—detects a coronary blockage, we assume the physician will treat the blockage, per our model above. This reflects clinical practice, in which a positive test is synonymous with treatment, and a negative test—i.e., one that does not change the treatment course—is a wasted test.[20] We thus define our treatment variable S_{ij} over the same 10-day window after ED visit j used for testing, as an indicator for the presence of procedure codes indicating stenting or coronary artery bypass graft surgery (CABG) for patient i.[21] This measurement strategy follows naturally from the goal we hope to achieve: to improve the targeting of tests, under current standards of medical knowledge and practice, and current performance of testing technology.[22] Table 2 shows that, among the tested, 12.9% received stents and 1.8% had a coronary artery bypass surgery (CABG) in the 10 days after their ED visit.

One reason this assumption could be problematic is because physicians may over-treat: that is, they deliver treatment S even when B is not present. This is particularly common in settings outside of the ED, where physicians often refer their clinic patients for elective stress testing or catheterization because of longstanding symptoms, for example, of chest pain or shortness of breath. These tests can reveal 'stable coronary disease,' and lead to stenting. But because these stents do not target new blockages, their benefit has been questioned in light of emerging evidence (Al-Lamee et al., 2018). We emphasize, however, that this is very different from the setting we study: patients coming to the ED for a new problem. Here, by contrast, there is incontrovertible evidence of large treatment benefits dating back to the 1980s (Amsterdam et al., 2014). That said, we account for the possibility that physicians over-treat in the form of the treatment effect parameter we use in our analyses (and explore a range of reasonable estimates in Appendix). Treatment effects reported in clinical trials should 'price in' over-treatment of patients who do not benefit: the more patients who are treated but do not truly have B, the lower the average treatment effect in the trial. So we perform sensitivity analyses on our cost-effectiveness results using a range of treatment effects from different studies and clinical scenarios in the literature, and our preferred parameter takes estimates at the bottom of the reasonable range (reviewed in Amsterdam et al. (2014) and Bavry, Kumbhani, Rassi, Bhatt, and Askari (2006)).

Beyond over-treatment, we are not aware of evidence of other variation in treatment

[20] We do not account for the intrinsic value of information, i.e., of 'knowing' that one does not have a blockage. To the extent that a negative test for heart attack triggers other tests, these other tests would need to be evaluated on their own merits.

[21] CABG is typically done if catheterization identifies multiple severe blockages that are not amenable to stenting. Thus the sequence of decisions leading to CABG is similar to stenting: it effectively requires diagnostic catheterization, even though stents are not delivered, to map out the coronary anatomy and assess eligibility for surgery. As with the differences between stress tests and catheterization, we consider the benefit to be similar, but keep track of which patient had which treatment in our cost-benefit analysis.

[22] As above, we do not assume tests detect all blockages: if B is present, we will only observe treatment S with probability q. This means that, in our model of testing, all tested patients pay the cost of testing c_T, but only the $(1-q)$ patients who test positive benefit from treatment (and pay costs of treatment c_S). This diminishes the value of testing.

conditional on test results. Test results for catheterization are based on objective measure of blood flow through the coronary arteries, which either visualizes a blockage or not; based on these results, either a stent is placed in the blockage or not. So while there is ample evidence of bias in testing decisions—for example, doctors are less likely to refer patients for testing for heart attack when presented with vignettes accompanied by randomly assigned pictures of women and minorities (Schulman et al., 1999)—bias in treatment, conditional on test results, is likely to be less widespread (though not impossible). A physician would need to overrule the objective measures, and clinical guidelines, to deviate. There is no clear reason to believe that there are widespread types of biases that would allow testing, but discourage treatment in tested patients. Finally, we might worry about correlation between the testing threshold and the treatment decision. This is mitigated somewhat by the fact that the physician referring the patient for testing (the emergency physician) and the physician performing the test and treatment procedure (the cardiologist) are two different people.

2.3.4 Undiagnosed Blockage in the Untested

Our measure of adverse events resulting from untreated (and undiagnosed) heart attack is drawn from the clinical literature. Because effective treatment for heart attack emerged only in the early 1980s, there is a large body of fairly recent research documenting the fate of patients with untreated heart attack.[23] Clinical trials of diagnostic and treatment interventions for heart attack (e.g., CT-angiography to diagnose coronary disease, e.g., Litt et al. (2012); or statins to treat high lipids, e.g., Ridker et al. (2008)) commonly use a basket of events derived from this literature, 'major adverse cardiac events,' as their primary outcome. We replicate this in our data, using methods similar to observational clinical studies (e.g., for decision rules: Than et al. (2011), Poldervaart et al. (2017), and Sharp, Broder, and Sun (2018)) that have shown excellent agreement with expert judgment after chart review (e.g., Wei et al. (2014)).

Specifically, we form an indicator $A_{ij} = 1$ if we observe adverse events for patient i, over the 30 days after visit j when complications from heart attack peak, in any of three categories. First, the blockage could worsen, prompting the patient to return for a delayed diagnosis of heart attack (some of which lead to treatment, in which case we would observe $S = 1$; given the delay between onset and diagnosis, however, treatment has been shown in clinical trials to be less valuable). To capture these missed opportunities, the literature generally relies on diagnosis codes for heart attack. But diagnosis codes are largely generated for billing purposes, meaning there are incentives to 'up-code' visits to support increased reimbursement. To deal with this, we introduce an additional criterion relative to the literature: when we see diagnostic codes for heart attack, we confirm them by looking at quantitative results from a concurrent laboratory test, troponin, that measure the extent of damage to heart muscle. This works as a check on spurious or erroneous coding, and provides an objective way to confirm the nature and severity of heart attack. (Laboratory results are not

[23]For example, trials comparing home vs. hospital management of heart attack in the late 1970s, which commonly showed no benefit to in-hospital treatment, tracked a range of clinical outcomes in diagnosed but untreated patients (Mather et al., 1976; Hill, Hampton, and Mitchell, 1978).

present in insurance claims data, so we do this only in the main hospital record dataset.) Second, some patients experience cardiac arrest, from the arrythmias precipitated by heart attack. We would see this in the form of diagnosis codes, or alternatively for procedure codes indicating cardiopulmonary resuscitation (CPR). Third, because arrythmia can strike suddenly, before the patient can reach the hospital in time to be diagnosed or treated, the patient might simply drop dead—often outside of the hospital. This normally poses a major problem for observational studies, because out-of-hospital deaths are not recorded in hospital records. To deal with this, we link hospital records to state Social Security data. Together, these data allow us to form A_{ij}, our proxy for blockage in the untested.

To use the rate of these adverse events as a measure of under-testing, we must compare it to some upper bound that is widely accepted by clinicians. Fortunately, such bounds are common in the clinical literature on decision rules, in particular those that seek to help doctors test for heart attack (e.g., TIMI: Antman et al. (2000), GRACE: Tang, Wong, and Herbison (2007), HEART: Backus et al. (2010) and subsequent validation studies, e.g., Than et al. (2011), Poldervaart et al. (2017), and Sharp, Broder, and Sun (2018)), as well as studies of new diagnostic technologies (e.g., CT-angiography: Litt et al. (2012)) or guidelines for preventative treatment (e.g., with statins: Ridker et al. (2008)) of heart attack. All these studies share the need to define a maximum allowable rate of adverse events in untested (and thus untreated) patients. Practically, if a group of patients is ex post found to have an adverse event rate of over some bound θ, they determine that this group should have been tested or treated. This line of research guides routine testing and management decisions in clinics and hospitals, and underlies recommendations from professional societies. It thus gives us objective thresholds for levels of risk, over and above the base rate, that would mandate testing. While some details of studies vary (e.g., some use follow-up periods after ED visits of 30 vs. 60 days), we choose the most conservative parameters and define under-testing as adverse event rates greater than $\theta = 0.02$ over the 30-day window after visits. Importantly, we do not assume these thresholds are optimal, we do assume that physicians believe them to be optimal, and thus would not knowingly leave high-risk patients untested, by the standards widely used in clinical research and practice.

Because these bounds are defined in terms of the average rate of adverse events in undiagnosed and untreated patients, it implicitly takes account of the base rate of adverse events (i.e., that $Pr(A_{ij} = 1|B_{ij} = 0) > 0$). But it does not account for two other factors that affect measurement of A_{ij}. First, as laid out in our model above, not all adverse events resulting from untreated heart attack will manifest in the first 30 days after heart attack (i.e., $Pr(A_{ij} = 1|B_{ij} = 1, S_{ij} = 0) < 1$). So in some analyses we consider longer-term outcomes A^{ℓ}_{ij}, where we measure adverse event over the entire year after visits. While this is more complete, even this measure will not capture downstream rates of heart failure or delayed arrythmias, which persist over the lifetime of patients affected. Finally, apart from mortality which we ascertain using Social Security data, all other events are only measured if the patient returns to the same health system we study for care. If the patient instead returns to the hospital across the street—in our setting, for example, there is an large tertiary care hospital 2 blocks away from the large tertiary care hospital we study, that is unaffiliated—we

will not observe it. On net, these issues mean we take our measured A_{ij} to be a lower-bound on the true number of adverse events in the population we study.

Table 2 shows data on adverse events.[24] Tested patients had worse outcomes on average, with for example 1.7% dying vs. 0.4% in the untested. The rate of adverse events also appears higher, but recall that any intervention performed in the wake of testing meets the clinical definition of adverse event (need for treatment of heart attack), so these rates will be mechanically higher in tested patients.

2.3.5 Comparison of Tested and Untested Outcomes

So far, we have defined entirely separate measures of blockage in the tested and untested sets. This is natural because what we observe in the short period after visits will differ greatly depending on whether a patient is diagnosed and treated, or not. But for a variety of reasons, we also wish to compare tested and untested patients directly. To do so, we create a longer-term measure of adverse events A^{ℓ}_{ij}: starting at 30 days after the visit j, we set $A^{\ell}_{ij} = 1$ if we observe any adverse event until 365 days after the visit. Starting measurement at 30 days gives us a marker of of adverse events that is formed similarly in both groups: if we included the short-term period after visits, when the measurement process for adverse events is very different in the tested vs. untested, we would mechanically observe higher adverse event rates in the tested. To see the problem, imagine two patients in the ED with a coronary blockage, one tested (and treated) and one untested. In the tested set, we will see a flurry of activity in their medical records over the next 30 days: the tests, but also higher rates of heart attack diagnoses and procedures related to heart attack resulting from the test, as well as similar diagnostic codes occurring for follow up visits related to the event. All this simply reflects close observation, both in the hospital and in the immediate aftermath of the visit, to ensure they are safe—not new adverse events. In the untested set, by contrast, there is no such level of scrutiny, but rather only adverse events in initially untested and undiagnosed patients. This makes comparisons in the first 30 days potentially misleading.[25] After this period, however, we can compare the rates of adverse events with more confidence. Table 2 shows that tested patients indeed do go on to have higher rates of long-term adverse events, at a rate of 6.4%, vs. 3.0% in the untested, reflecting their higher risk. But this is more reasonable than the comparison of "adverse events" in the 30 days after visits, where tested patients have a rate of 26.1%—simply reflecting diagnoses and procedures in the wake of the test itself—vs. 1.2% in the untested.

[24] Rates in untested patients exclude those who had evidence of heart attack in the ER, described in more detail below.

[25] The exception here is death: because we measure this via linkage to Social Security, it is measured similarly in both groups. So we perform sensitivity analyses using death only over the entire year, including the first 30 days.

2.3.6 Eligibility for Testing and Treatment in the Untested

Tested patients are, naturally, eligible for testing, so we know that $K_{ij} = 0$. But untested patients may have been ineligible, for example because they would be harmed by the invasive nature of testing or treatment (which both involve catheterization, directly or indirectly after a positive stress test). In such patients, if we observe $A_{ij} = 1$, we cannot simply conclude the physician under-tested. The adverse events may have been inevitable, rather than the result of a failure to test and treat in the ED. For this reason, our main sample excludes some patients with general poor health or frailty (age over 80 years old, advanced cancer or dementia, palliative care, etc.). But physicians might see other elements of K that we do not observe.

Electronic health records are very helpful here, because they let us observe granular data elements from ER records that indicate physician awareness of risk. In untested patients with these markers of risk, we can infer that physicians likely considered the possibility of heart attack, but decided not to test deliberately. First, we observe the diagnosis that the physician assigns at the end of the ER visit. In some untested patients (2.97%, or a set as big as the entire tested set), the physician explicitly notes a diagnosis of heart attack in the ER.[26] In these patients, where the physician notes a diagnosis of heart attack but nonetheless decides not to test, we can infer that she has decided they would not be candidates for invasive treatment.[27] Second, we observe the results of troponin testing in the ED. Troponin is a laboratory study, often done along with other basic tests, that physicians routinely use to screen for heart attack (in our sample 13.9% overall have a troponin test done). It is a protein found in heart muscle cells that is released into the blood stream when these cells die and rupture. While a positive troponin is not definitive proof of heart attack—small elevations can have various causes besides heart attack (Jeremias and Gibson, 2005)—if the physician sees a positive result, it is safe to assume that she was aware of the possibility of heart attack. So in the 6.88% of untested patients with a positive troponin (2.4 times the size of the tested set), we can infer that she decided the patient was not eligible for testing.[28]

The ability to observe ECG and troponin data also give us insight into another important set of patients: those whom the physician does *not* suspect of heart attack. While patients can have ECGs and troponins ordered for reasons other than heart attack, if we see a patient without these tests ordered, it is highly unlikely that the emergency physician suspected heart attack. Table 2 shows that 27.7% of untested patients had ECGS done in the ED, and 9.2% had troponin tests. [29]

[26] There are two good reasons physicians do this: first, for medical communication, and secondly for billing, since ER visits are reimbursed proportional to patient complexity.

[27] We confirmed this intuition via a hand-review of a small sample of charts. Reasons included other severe illness, patient or family preference, previously known severe coronary disease that has proven refractory to treatment, and other reasons.

[28] Appendix 6.1 contains data on testing decisions, yield of testing, and adverse outcomes as a function of patients' troponin results and whether or not they had an ECG in the ED. Among tested patients, markers of risk are linked to higher test yield, as we would expect.

[29] ECGs are done in the course of tests for coronary blockages as well as in the ED, so all the tested have an ECG performed.

2.4 Algorithm Design

Most risk prediction tools for heart attack in the medical literature use a handful of clinical variables as predictors, for example elements of the medical history, certain laboratory studies, or interpreted features of the electrocardiogram (e.g., TIMI, GRACE, or HEART scores). As noted above, claims or electronic health records, by contrast, contain a vast set of other potential predictors that are increasingly being used as inputs to machine learning models in medicine.[30] Building on this work, we design a machine learning algorithm to accurately predict risk out-of-sample using a wealth of EHR data. While we cannot share patient-level information to protect privacy, our entire code repository is publicly available on GitLab.

2.4.1 Predictors

To form the inputs to our predictive model, we begin by transforming the discrete person-date data described above (e.g., a cholesterol value of 200mg/dL, or a hospitalization for heart failure, on a given day before the ED visit) into summary statistics (counts, averages, standard deviations, etc.) over discrete time periods (i.e., 0-1 months, 1-12 months, and 12-24 months prior to a visit). For diagnosis and procedure codes, as well as medications, we additionally take advantage of the fact that these codes are nested in categories: so we aggregate them into clinically meaningful 'super-variables,' by collapsing at the level of hierarchical taxonomies defined by the Agency for Healthcare Research and Quality's Clinical Classification Software (with minor modifications, available on our online code repository), and the ATC classification for medications.[31] This results in one variable for each time period, describing occurrences over a short medium and long windows before a given visit, and for each semantically-grouped diagnosis, procedure, or medication group. We dropped variables missing in over 99% of the training set, leaving a vector X_{ij} of 16,381 predictors for visit j by patient i.

We were very careful to form these variables so that the information they contain was uniformly available to the physician at the time of the decision. This is harder than it seems: for example, sometimes physician notes are dated on the day of the ED visit, but are completed by the physician days or even weeks later—after information on the results of testing become available. The data available during the course of the ED visit, like the results of laboratory testing or the electrocardiogram, are likewise 'downstream' from the decision making process we aim to assess: only patients suspected of a heart problem will

[30] Rajkomar, Dean, and Kohane (2019) provide a helpful review of recent work. Notable examples include Ghassemi et al. (2014), Rajkomar et al. (2018), and Henry, Hager, Pronovost, and Saria (2015), who use these tools to predict clinical outcomes, and Miotto, Li, Kidd, and Dudley (2016) who predict a variety of future diagnoses. We are necessarily brief in our description of machine learning methods; see Mullainathan and Spiess (2016) or Athey and Imbens (2019) for a more thorough overview with references.

[31] As an example, the occurrence of a low-level diagnosis or procedure code (e.g., E018.2: Injury from activities involving string instrument playing) 100 days before a patient's visit would be aggregated into a broader clinically meaningful categories (e.g., E000-E999: External Causes Of Injury) over a specific time period (i.e., 31-365 days prior to visit).

have certain test results present—but we wish to create predictions irrespective of whether the physician suspected a heart problem. So we stop incorporating any information from the EHR starting the moment the patient arrives at the ED triage desk. Later, we will use some data from the ED visit itself to infer the physician's level of suspicion for heart attack, but we do not include any of these data into the predictive model.

2.4.2 Training Procedure

Our goal is to form estimator $\widehat{m}(\cdot)$ of the risk of blockage $Pr(B_{ij} = 1)$, on the basis of observed covariates X_{ij}. As laid out in our model above, B_{ij} is not observed directly, but our dataset gives us two ways to measure it: a positive test result leading to treatment S_{ij} when $T_{ij} = 1$, and an adverse event A_{ij} when $T_{ij} = 0$.

Because machine learning models over-fit to the data on which they are trained, we ensure that our predictions are valid out-of-sample by randomly splitting the sample into a training set for model development, and a hold-out set for model validation. If patient i has more than one ED visit j in our sample, we can have several observations on the same patient. Because these visits happen at different times, both the outcome of the visit (e.g., was the patient tested) and the background variables we observe about the patient (e.g., their most recent blood pressure) vary—but of course they are not independent. Practically, we handle this by splitting our dataset at the patient level, rather than the observation level, so that all visits from a given patient are assigned exclusively to either the training or hold-out set. We also split out a small 5% 'ensembling set' from the training set (and distinct from the hold-out set), which we use to calibrate our ensemble. This means observations (and patients) fall into three mutually exclusive sets: training (70%), ensembling (5%), and hold-out (25%). The estimator is trained on the first two sets, and all results are shown exclusively in the hold-out (except where noted explicitly).

In the training set, we form four individual estimators, two in the tested and two in the untested, that will later be combined into an 'ensemble' estimator $\widehat{m}(X)$. First, in the tested patients, we fit two distinct machine learning models, gradient boosted trees and LASSO, both designed to handle large sets of correlated predictors (Friedman, 2001) to predict treatment $S_{ij} = 1$ using observed covariates X_{ij}. This results in two estimators that predict treatment in the tested, one gradient boosted tree and one LASSO.[32] In the untested patients, we fit two similar models to predict $A_{ij} = 1$ using X_{ij}. We do so because, as noted above, we would expect there to be signal for predicting B_{ij} in both treatment S_{ij} when $T_{ij} = 1$ and adverse event A_{ij} when $T_{ij} = 0$. So a model that was useful for predicting one would have signal for predicting the other. This also let us take advantage of the far larger sample size in the untested. Using each of these four functions (one LASSO and one gradient boosted tree, for each outcome), we generate four predictions for each observation in the 5% ensembling set. Because we view S as a more reliable proxy than A, we design the ensemble estimator to predict $Pr(S_{ij} = 1)$: in the ensembling set, we fit a no-intercept

[32] A trivial way to see the benefit of machine learning methods here is that we have only 5,755 tested patients in our train set, and 16,381 predictors, so traditional functional forms are not possible.

logistic regression to predict $S_{ij} = 1$, using the four predictions on the right hand side. The output of this weighted combination forms the final ensemble model $\widehat{m}(X)$.

We also note another, related estimator we form in the training set. For the subset of patients who had ECGs done in the ED, we obtain the ECG waveform and use a convolutional neural network to form similar predictions on S_{ij} when $T_{ij} = 1$. This model, which we denote $\widehat{m}_{\mathsf{ECG}}(X)$, described in more detail in Appendix 3.

2.4.3 Evaluation procedure

Our empirical tests happen in the holdout set $\mathcal{H}$ of n=62,850 visits by 32,791 patients. We emphasize that the model was never exposed to these hold-out data in the training process. Of these, the tested subset $\mathcal{T}$ is comprised of the 1,834 visits (1,528 patients) for which $T_{ij} = 1$. In visits for which $T_{ij} = 0$, we first exclude visits where $K_{ij} = 1$ (on the basis of diagnosed heart attack or positive troponin test). We assign the remaining 59,987 visits by 31,263 patients to the untested set $\mathcal{U}$. For our bounding exercise, we further subset these patient-visits to form $\mathcal{U}_{K^-}$, the set of visits that lack an ECG or a troponin result.

We then apply $\widehat{m}(X)$ to form out-of-sample predictions on S for all visits in $\mathcal{H}$, which we interpret as the probability of diagnosed blockage when tested, based on observable factors in the tested. We use these predictions to form the following statistics. First, to quantify over-testing, we look in the set of tested patient-visits $\mathcal{T}$ for patients in a given bin of predicted risk, $\widehat{m}(X_{ij}) \in r$. We define over-testing if the average yield of testing $\bar{S}_r$ falls below the value threshold defined above:

$$\bar{S}_r = \frac{\sum_{j \in \mathcal{T}_r} S_{ij}}{|\mathcal{T}_r|} - \frac{c_T}{q(\tau - c_S)}$$

To account for multiple visits by the same patient, we cluster standard errors at the patient level in all analyses.

In addition to the rate of treatment $\bar{S}_r$ among tested patients in a given risk bin, we can directly calculate the value of testing $V_m(r, T = 1)$ as above. The inputs are whether a given patient has a blockage B_{ij}, proxied by treatment S_{ij}, as well as three additional parameters: the benefit of treatment τ_{ij}, and the costs of testing and treatment c_{Tij}, c_{Sij}. Since our estimates are formed in the tested, all $T_{ij} = 1$ and all patients pay some cost of testing. But the specific financial costs differ by testing strategy, and we allow them to vary using estimates from the literature specific to the type of test the patient has had (non-invasive stress tests vs. diagnostic catheterization or both). We also account for the health costs of testing for those receiving catheterization (the only one that materially affected estimates is related to the rare but catastrophic procedural complication of stroke). The sum of these forms c_{Tij}. For the subset of patients for whom $S_{ij} = 1$, we assign the costs and benefits of treatment for heart attack. To estimate the financial costs of treatment, we similarly use estimates specific to the type of treatment the patient has received (stenting vs. CABG). To estimate its benefits, we use the approach from the clinical cost-effectiveness literature (summarized in Mahoney et al. (2002)). We first estimate η_{ij}, in terms of how many life-years a given patient would lose from a heart attack at the time of their visit, depending on

their age and comorbidities. This includes both fatal and non-fatal cases, the latter using a standard discount rate for quality of life decrements due to heart attack, applied to all remaining life years. We then apply estimates from clinical trials to estimate the reductions in these losses we could expect from timely treatment $\bar{\tau}$.[33] The product $\eta_{ij}\bar{\tau}$—the number of life years saved by treatment—forms our estimate of τ_{ij}. This lets us calculate the cost per quality adjusted life year in each bin of predicted risk in the tested[34]:

$$\bar{V}_r = \frac{\sum_{j \in \mathcal{T}_r}(c_{Tij} + S_{ij}c_{Sij})}{\sum_{j \in \mathcal{T}_r} S_{ij}\tau_{ij}}$$

Second, to quantify under-testing, we look in the set of untested, eligible patient-visits $\mathcal{U}$ in a given bin of predicted risk $\widehat{m}(X_{ij}) \in r$. We first form 'naive' estimates of the yield of testing in the untested, analogous to $\bar{S}_r$ in the tested:

$$= \frac{\sum_{j \in \mathcal{U}_r} \widehat{m}(X_{ij})}{|\mathcal{U}_r|} - \theta$$

By simply imputing the predicted yield of untested patients, conditional on X and using estimator $\widehat{m}(\cdot)$ formed in the tested, we have formed an estimate of under-testing that mirrors those used in the literature: a physician failing to test when the statistical model indicates it is worth testing. Based on this, we can also generate estimates of $\tilde{V}_m(r, T = 0)$, by modeling the relationship between $\widehat{m}(X_{ij}$ and $V_m(r, T = 1)$, and using this to generate estimates for $V_m(r, T = 0)$ on the basis of $\widehat{m}(X_{ij}$.

Our approach, by contrast, does not take these estimates at face value. Rather, we use predictions as only a starting point for a more empirical definition of under-testing. Consider the set of all observations our naive estimator flags as high-yield based on $\tilde{S}_r$ (or cost-effective based on $\tilde{V}_m(r, T = 0)$). For this set, we further estimate $Pr(A = 1|\widehat{m}(X) \in r, T = 0, K = 0)$ by calculating the realized rate of adverse events $\bar{A}_r$, and comparing it to θ, in each bin of predicted risk r:

$$\bar{A}_r = \frac{\sum_{j \in \mathcal{U}_r} A_{ij}}{|\mathcal{U}_r|} - \theta$$

In addition to providing an empirical verification of model-predicted under-testing, this also allows us to measure how often patients the naive estimator considers cost-effective actually go on to have an adverse event. Because of the importance of properly measuring the $K = 0$ condition, in order to conclusively establish physician error, we conduct a sensitivity analysis to lower-bound the estimate of $\bar{A}_r$, by recalculating these statistics for set $\mathcal{U}_{K^-}$ of visits without ECGs or troponins. In these patients, physicians are unlikely to have suspected

[33] We assume a constant treatment effect in all these patients, because we have excluded those for whom $K = 0$. We choose a conservative estimate of $\bar{\tau}$ (a 25% reduction in mortality and morbidity), and report results from a range of plausible estimates below. A full accounting of individual costs, benefits, and assumptions is in Appendix 4.

[34] To avoid dividing by zero in bins with few treatments, we use larger bins of $\widehat{m}(X_{ij})$ to present results for $\bar{V}_r$ (based on risk quintiles in the tested) than for the analysis of yield $\bar{S}_r$ (based on deciles).

heart attack, let alone suspected it strongly then ruled out invasive treatment because of contraindications.

Our final piece of evidence on under-testing comes from a 'natural experiment,' described in more detail below, that lets us directly estimate the effect of testing on health. This is perhaps the best way to establish whether physicians should be testing high-risk patients. Depending on the triage team working when a patient shows up to the ED, her probability of being tested varies in a way unrelated to underlying risk. By comparing rates of longer-term adverse events A_r^ℓ in this setting (which unlike S_{ij} and A_{ij} are measured similarly between tested and untested), we can compare real health outcomes in patients 'assigned' to more or less testing. We are particularly interested in distinguishing between the average effect of testing on all patients, and the effect on patients at a given level of predicted risk. Detecting a health effect of testing in this setting would indicate that physicians are missing opportunities to test in ways that would benefit patients.

2.5 National Medicare claims

To check the generality of the results we see in our single-hospital dataset, we replicate them in a nationally-representative 20% sample of Medicare claims data. (As is usual, we exclude non-fee-for-service patients, since we do not observe their full claims history.) This comprises 20,059,154 ED visits over a four and a half year period from January 2009 through June 2013. Applying similar exclusions to those used in the EHR data, we arrive at a final sample of 4,425,247 Medicare visits by 1,602,501 patients. In 195,287 visits, patients were tested, and in the remaining 4,229,960 they were not. Of the tested, 124,736 had stress tests and 84,481 had cardiac catheterization in the 10 days after visits; 24,126 received treatments (stents, CABG). We fit the algorithm as above, with two exceptions driven by data limitations in insurance claims. First, Medicare claims contain only the first two categories of data above (demographics, including mortality, as well as diagnosis and procedure codes), meaning three important types of data present in EHR data are absent: (i) biomarkers, meaning we are unable to confirm diagnosis and procedure codes in our definition of adverse events, and ECGs which are not well measured in claims; and (ii) other laboratory studies and vital signs are likewise absent and cannot be used as predictors; (iii) triage chief complaint, timing, and presence of an ECG are also not recorded in claims. Second, since we do not observe granular time stamps in billing data, we are more cautious in fitting our predictor, and only form variables X up until 3 days prior to ED visits, to ensure that no information from the visit leaks into the predictor set.

3 Errors in Testing Decisions

3.1 Over-testing

We begin by calculating the average yield of testing, as is common in the health policy literature. In our hold-out set, the fraction of tested patients who go on to have a positive

test and be treated, $\bar{S} = 0.145$. We then examine how yield and cost-effectiveness varies with algorithmic risk predictions, $\bar{S}_r$. Figure 1 shows the results graphically. In the hold-out set of patients, we rank-order patients by their predicted risk on the x-axis. To calculate the realized yield of testing among tested patients, we bin them into risk deciles, which we define using bin cutoffs $r = [\underline{r}_T, \bar{r}_T]$. Thus each point shows on the y-axis the yield for patients in a given risk bin; exact numbers are shown in Table 3. As the algorithmic risk prediction increases, realized yield in the hold-out set increases. Moreover, among the patients doctors chose to test, we see not only large variation in predicted yield on the x-axis, but also in actual realized yield on the y-axis. The relationship is monotonic, and the model is able to identify large groups of patients with very low yield amongst the tested. Indeed, the lowest-risk bin of tested patients had only a 1.6% yield, or one-ninth the average rate.

3.1.1 Cost-effectiveness

Using the methodology described above, we can also calculate a related quantity, the average cost-effectiveness of testing, $\bar{V} = \$86,68$ per life year. Accepted thresholds for judging cost effectiveness typically range from $100-150,000 in the US (Neumann, Cohen, and Weinstein, 2014). So testing as a whole in this setting could be considered cost-effective, at typical life-year valuations. (Comparing our results to the literature on stress testing, we find higher yield and far more favorable cost-effectiveness, likely because we consider all tests—catheterization as well as stress tests—rather than stress-tests alone.)

The bottom panel of Figure 1 shows variation in cost-effectiveness of these tests as a function of predicted risk, $\bar{V}_r$. We again rank-order patients by their predicted risk on the x-axis, this time in quintiles.[35] The y-axis shows, for a given risk quintile, the cost-effectiveness of testing these patients, in units of cost per quality-adjusted life year. This reveals that a substantial fraction of patients physicians choose to test are very low-value: for example, in our preferred specification, the lowest-risk quintile come at a cost of $1,183,936 per life year. For comparison, biologics for rare diseases (like gene therapy, 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.

We might worry that this estimate is sensitive to the particular choice of parameters for the cost-effectiveness analysis. In particular, while the costs of testing are well-defined, there could be a range of plausible estimates for the amount of benefit patients get from prompt diagnosis and treatment of heart attack. Our preferred specification uses a 25% reduction in mortality and morbidity from heart attack, which we view as very conservative in light of the ranges reported in the literature (and detailed in Appendix 4: Table 6). Figure 1 shows how using estimates from a 20% to 30% reduction affects our cost-effectiveness, and the Appendix contains additional estimates for still wider ranges. Even at the top of this range of treatment benefits, cost-effectiveness is highly unfavorable compared to the usual thresholds for willingness to pay in the lowest-risk 40% of tests. This highlights the extreme

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

low-value of testing on the margin, and is particularly striking when compared to the average yield, which made testing appear quite cost-effective on average, hiding the extent of waste.

Viewing individual tests through the lens of predicted risk also allows us to simulate the effects of dropping specific marginal tests, at a given life year valuation. This counterfactual depends on the assumption that dropping a negative test would have no health consequences for the patient (besides saving her the health costs of the test). For example, at a \$150,000 life-year valuation, we would drop 49.1% of the lowest-value tests, with a combined cost of \$10,284,000; this would have foregone 11.3% of all treatments in the population, which would have generated a \$3,699,845 value in life years. In other words, the relevant policy prescriptions are no longer restricted to decisions about testing as a whole, but can instead accommodate tailored, prospective policy guidance at the level of the individual patient.

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, we also explore other counterfactuals in Appendix 5. We find that, as predicted by our model, physicians tend to begin with stress testing in lower-risk patients, presumably because the costs of these tests are lower. However, despite its lower cost, stress testing alone is still far lower-value than catheterization. When we calculate cost-effectiveness, we see that stress tests find far too few blockages to be worthwhile, even accounting for their lower costs, while the value of catheterization is more in line with the overall results above. This provides support to previous arguments that the current practice of stress testing is so low-value that eliminating them altogether would improve welfare (Prasad, Cheung, and Cifu, 2012). Of course this does not imply that high-risk patients should not be tested at all: similarly high-risk patients tested with catheterization have very cost-effective yields. So to simulate a counterfactual where we bypass stress tests altogether and go straight to catheterization for all patients, we create an upper bound of cost-effectiveness. We optimistically assume we can cut all stress tests and still, without the benefit of that test, identify the patients who went on to a positive catheterization. This of course increases the value of testing high-risk patients, but even under these conditions, the implied cost-effectiveness in the lowest-risk 20% of tests (catheterization) remains very low, at \$434,800 per life-year.

Figure 1 also shows the very high value of testing for the highest-risk patients. In the highest-risk quintile, for example, the cost per life year is only \$48,831 per life year, well within bounds of willingness to pay and comparable to widespread interventions like dialysis for kidney failure. Under our preferred specification, a substantial fraction of the tested, 50.9%, would be considered cost-effective at the \$150,000 per life year threshold. Here again, as above, our main results do not appear to be particular sensitive to the choice of treatment benefit parameter for cost-effectiveness. Even taking a very conservative lower-bound of 10% reduction in mortality and morbidity from prompt diagnosis and treatment of heart attack, these tests generate a value of \$61,473.

3.2 Under-testing: Model Predictions

We now turn to results for the untested, again in the hold-out set of patients, but now removing all those who are ineligible for treatment or testing (i.e., those with $K_{ij} = 1$). In columns (1) and (2) of Table 3, we summarize the yield and the cost-effectiveness of testing, by predicted risk (i.e., quintiles of $V_m(r, T = 1)$, calculated in the set of tested patients). Column (3) then considers the testing rate among all patients (tested and untested) in a given risk bin (i.e., $Pr(T = 1|\widehat{m}(X) \in r)$). Risk bins r are formed using risk deciles in the tested for columns (1) and (2) as above, and using risk deciles formed in the full holdout for column (3). Two patterns emerge: first, doctors are more likely to test higher risk patients. But second, if we consider the highest-risk bin, who appear to be highly cost-effective, only 54.6% are actually tested. As another point of comparison, if we take the riskiest 2.97% of patients in the entire hold-out set, a group chosen to be the same size as the entire tested set, the testing rate is only 26.5%. The model predicts that this last set in particular is very high-risk: if we simply calculate mean predicted risk in these patients (i.e., $\tilde{S}_r$), we get 25.1%, compared to a mean predicted risk of 14.3% across the entire tested set.[36]

This raises the possibility that, in addition to over-testing, doctors may also be under-testing. The existence of over- and under-use in health care has been raised in several recent papers. Most closely related, Abaluck, Agha, Kabrhel, Raja, and Venkatesh (2016), build a structural model of risk for pulmonary embolism. They find that physicians vary in where they set their risk thresholds for testing, leading some physicians to test marginal patients with extremely low absolute risks. In addition to this clear evidence of over-testing, they also find that physicians mis-weight observable factors correlated with high risk, such that high-risk patients are often left untested. The implication is that, in a counterfactual world where these apparently high-risk patients were tested, we would have found positive test results, and concluded that this was under-testing. There is substantial support for this view in the medical literature on diagnostic error, which traces back adverse health events and malpractice claims to physicians' failure to test high-risk patients (Kohn, Corrigan, and Donaldson, 2000; Graber, Franklin, and Gordon, 2005; Newman-Toker, Moy, Valente, Coffey, and Hines, 2014; Singh, 2013).

But the evidence on under-testing from the literature, like our own findings so far, is at best suggestive: a statistical model that disagrees with the physician's decision to test or treat a given patients. In order to convincingly document under-testing, though, a basic econometric problem must be solved. We do not observe test results for untested patients. It is one thing to assert under-testing on the basis of a structural model, thus relying on imputation of outcomes we do not actually observe. It is quite another to show it empirically. This is particularly true in settings where physicians have a considerable information

[36]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 10.1, 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.

advantage over the statistical model. Given the wealth of private information they see and we do not, it is very possible physicians are leaving these patients untested for good reason.

To illustrate the magnitude of this problem, we first calculate $\tilde{S}$, the yield the model would predict in untested patients at a given level of predicted risk. We find that the untested as a whole have a predicted yield of 4.1%. By comparison the realized yield in the tested is 14.5%. However, because of the far larger size of the untested set—the entire tested set would make up only 3.1% of the untested set—there are many more patients with very high predicted risk in the untested. For example, the model predicts that 2,738 untested patient-visits would have led to treatment, had they been tested, while in the tested set, physicians actually found only 266 patient-visits that led to treatment. This would imply that doctors are missing 90.6% of all heart attacks among patients passing through the ED. These back-of-the-envelope calculations suggest that model predictions may be missing important signals, and over-predicting risk.

3.2.1 Capturing Risk Information from Electrocardiograms

To show this more precisely, we turn to a subset of patients in whom we have the opportunity to add an important source of physician private information: the electrocardiogram (ECG). The ECG is a fundamental part of how physicians forms their risk estimates on heart attack. Among the 29.8% of patients with ECG results in the ED, the physician can observe an important source of signal regarding the likelihood of blockage. So far, we have not included ECG data in our model, for a specific reason: not all patients have an ECG, and whether they have an ECG is itself a function of the physician's risk prediction. So it would be problematic to rely on these data for a general risk prediction function—it would not be applicable to all patients (i.e., we could not generate predictions for the 70.2% without ECGs). For the purposes of quantifying the physician's informational advantage, however, the ECG is very valuable. This is particularly true because ECG data are rarely included in statistical risk models: even if they were available for all patients in a cohort, they are housed in separate databases from standard EHR data and thus difficult to access. Further, they consist of waveforms that cannot be accommodated easily in traditional statistical risk models. So our results translate more broadly into a range of prediction models in the literature that do not incorporate such important sources of signal available to physicians.

For those patients with ECG data available, we first explore how physicians appear to be using obvious features of the ECG that are suspicious for heart attack, using the physician interpretation entered into the electronic record to accompany the waveform. We identify six important features of the ECG that arouse suspicion for heart attack, and use basic natural language processing to extract these from the free text report entered by the cardiologist to accompany the ECG in the health record. This lets us form a vector of indicators indexing these six findings (more details are in Appendix 3), as well an indicator for the cardiologist interpreting the ECG as basically normal (i.e., no indication of heart attack or other problems). We then run two parallel regressions exploring the relationship of these seven factors to both the testing decision, and the yield of testing, conditional on

predicted risk:

$$T_{ij} = \beta_0 + \beta_1 \widehat{m}(X_{ij}) + \beta_2 \mathsf{ECGFeatures}_{ij} + \epsilon_{ij} \quad (1)$$

$$S_{ij} = \gamma_0 + \gamma_1 \widehat{m}(X_{ij}) + \gamma_2 \mathsf{ECGFeatures}_{ij} + \epsilon_{ij} \quad (2)$$

We find that, of the six features suggestive of risk for heart attack, four are highly predictive of testing—and all four are additionally highly predictive of yield, with the same sign. In other words, physicians are making use of risk information in the ECG effectively. The feature of 'ST elevation,' for example, makes testing nearly 9 times more likely (5.3 p.p., SE: 2.0)—and more than triples the likelihood that testing will yield a positive result (15.5 p.p., SE: 5.5), conditional on predicted risk that does not incorporate the ECG signal. Having an entirely normal ECG, by contrast, makes testing 2.4 p.p. less likely (SE: 0.6), and reduces yield of testing by 5.1 p.p. (SE: 2.3), conditional on predicted risk. Full results are in Appendix 3.

While these individual features are clearly meaningful, both for the physician's decision making process and for the likelihood of blockage, we might wish to extract the maximum amount of risk information from the high-dimensional ECG signal, independent of the physician's interpretation. To do so, we build a new model of risk, $\widehat{m}_{\mathsf{ECG}}(X)$, that incorporates the structured risk information vector X_{ij} as well as the ECG waveform, using a convolutional neural net. It is otherwise trained and applied in the same way as our usual 'naive' risk predictor, that does not incorporate ECG information, $\widehat{m}(X)$. We then simply compare how much the addition of ECG risk information updates the original prediction, by subtracting $\widehat{m}_{\mathsf{ECG}}(X) - \widehat{m}(X_{ij})$. Across the entire population, we find that adding ECG information decreases risk for 97.6%, and increases it for 2.5%. In the highest-risk bin of untested patients (0.21%) based on $\widehat{m}(X_{ij})$, adding the ECG information decreases predicted risk for 100%;, resulting in 24.7% being dropped out of that highest-risk bin. It is revealing that predicted risk *on average* falls. Even if ECGs resulted in better prediction, why is the average prediction across the whole population changing so sizeably? The reason is intimately tied to the physician's information advantage. We are training models on the tested and (naively) extrapolating predictions to both tested and untested. If the physician is using unobservables to select who is tested, then the untested are less risky—even conditional on observables—than the tested. So a better predictor that uses these unobservables, will predict lower overall risk, which we see here.[37]

3.3 Under-testing: Evidence from Adverse Event Rates

We now examine whether there is under-testing. As noted earlier, we do not take the model's prediction at face-value. Instead, for untested high-risk patients we look for evidence that they had a heart attacks. In particular, we calculate the rate of adverse events in the 30 days after their visit. To put numbers on adverse events in context, we use clinical decision rules that provide thresholds for guidance on the level of realized adverse event

[37]In other words, the ECG results in better matching of predictions to reality: since most patients are negative, most patients are updated negatively, while the small number of positives are updated positively.

rates which merit testing; these correspond to the parameter θ in our model. As noted above, testing guidelines for heart attack commonly used in emergency rooms recommend considering testing for patients who would have a realized adverse event rate of 2% over the 30-day window after visits.[38] Importantly, since we use the 2% threshold to test for physician private information, we do not take a stance on whether it is physiologically optimal, only that it represents current physician understanding of who should be tested. If physicians are failing to test apparently high-risk patients because of private information, and applying the existing clinical knowledge, these high-risk patients should have rates below the thresholds. This is especially true given incentives to over-test.

Among all untested patients (excluding, as noted above, those with suspected or realized heart attack in the ER), the rate of adverse events in the 30 days after their emergency visit is 1.25%. Comfortingly, this is below the 2% threshold as one would expect: testing all untested patients is unlikely to make sense by this standard. We now focus on those patients predicted to be high risk by the algorithm; as in the model above, we display $Pr(A = 1|\widehat{m}(X) \in r, T = 0, K = 0)$ for ranges of predicted risk r in Figure 2. Untested patients are ranked by predicted risk and binned, with bins shown on the x-axis. The adverse event rate for each bin is then shown on the y-axis. For comparability, the Figure uses bin cutoffs for deciles formed in the tested, which means that bins are of unequal sizes in the untested: in particular, because the untested are lower-risk than the tested, bin sizes decrease in risk. Nonetheless, we see that the highest-risk patients have strikingly high rates of 30-day adverse events. For example, the highest-risk bin contains 0.22% of the untested, who go on to have an adverse event rate of 9.16%. The second highest-risk bin contains 0.78% of the untested and has adverse event rate of 5.97%; together the top two bins have an adverse event rate of 6.66%[to compare to no-ECG/Tn below]%. In fact, the cross over point for the 2% threshold is the 5^{th} risk bin which means that top 6 bins (which comprise of 8,522 people, or 14% of the untested ¡–i'd do percent of untested) all appear to be high-risk enough to merit consideration of testing. Together, these top 6 bins contain 8,522 untested and 1,100 tested people. As we said these risk bins are defined using deciles of the tested for comparability. Alternatively, if we were to form a set of the same size as the tested (the riskiest 1834 untested patients, or 3.1% of the untested), that group would have an adverse event rate of 5.18%. Together these results illustrate that in fact the untested with the highest predicted risk do go on to have very high adverse event rates, arguing against physician private information about risk.

We have placed considerable emphasis on the 2% adverse event threshold from the clinical literature for our definition of under-testing. While this is reasonable, given how widely such thresholds are used in clinical settings and in the health literature, it is very different from the approach we used to set a threshold grounded in cost-effectiveness for examining over-testing. Do these two align with each other? Since they are not in the same units – one is defined on adverse event rates while the other is on yield rates – they cannot be directly

[38]Recall that the clinical literature from which the threshold is drawn defines it to include the base rate of adverse events in the 30 days after visits. As another point of comparison, surveys of emergency doctors ask about their willingness to accept a heart attack miss rate. These find a tolerance of up to 1% for heart attacks in the ER (Than et al., 2013).

compared. Instead, to test for alignment, we form 20 bins of predicted risk in the hold-out set (based on risk ventiles in the tested) and calculate two quantities: (i) cost effectiveness in the tested; and (ii) realized adverse event rate in the untested (see Appendix 6). We can now for each bin examine how the two rates compare. In particular, we can see how bins at the threshold for one rate compare on the other rate. As we can see from Appendix Figure 2, the two thresholds align quite well. For example, if we picked risk bins we should test using the cost-effectiveness threshold of $150,000 per life year (as we did), then those same risk bins would imply we should test risk bins using an adverse event rate of 0.025— quite close to the 0.02 rate from clinical guidelines.

Another potential problem with a threshold for under-testing based on adverse events parallels a common critique of the clinical literature: its 'adverse events' are in fact discretionary utilization of care by physicians. Using diagnoses or procedures as a proxy for health outcomes can over-state benefits or harms, if these are measured with error due to incentives to over-test or treat (Welch, Schwartz, and Woloshin, 2011; Armstrong, 2018). For example, our adverse event measure includes subsequent diagnoses and treatments for heart attack or arrhythmia—but these depend on physicians' testing and treatment decisions, as well as hospital billing and coding practices. Thus an important question is the extent to which our adverse events are in fact driven by potentially discretionary care. We view this as unlikely for two reasons. First, as noted above, we do not rely on billing codes to determine heart attack alone. Rather, we additionally cross-reference these codes with laboratory evidence, and only classify an adverse events when we see a positive troponin test, indicating real damage to heart muscle. While this still requires a physician to decide to do a laboratory test, its result is a biomarker that cannot be manipulated. Second, we emphasize that nearly half of the adverse events we record in the highest-risk bin, 47.5%, were deaths, ascertained using Social Security data. Together, these two points provide reassurance that the adverse event rates we see constitute real health outcomes, not ambiguous proxy measures.

3.4 Failure to Account for Differences in Treatment Benefits

So far, we have relied on two methods to exclude untested patients with contraindications to testing or treatment $K_{ij} = 1$. First, we exclude all patients with evidence of poor health or limited treatment benefit, based on data before they set foot in the ED (patients over 80 year old, those with diagnoses like cancer, those receiving palliative care, etc.). Second, we exclude those in whom data from the ED visit indicates a suspicion of heart attack (either because the physician has assigned a diagnosis code of heart attack, or because we observe a troponin test that came back positive). The second step is particularly valuable, because it lets us exclude a broader group of cases where the physician was aware of heart attack in that particular visit, but decided not to test. However, because of the importance of carefully excluding ineligibles for our estimates of under-testing, we conduct two additional analyses to check the robustness of our measurement strategy for K_{ij}.

3.4.1 Patients With Clear Alternative Diagnoses

First, it is possible that in some cases a physician has some suspicion heart attack in the ED, but is not confident enough to document the diagnosis explicitly (or obtained a troponin test, but did not get a positive result, which is reassuring but not definitive). Importantly, in cases where such a serious condition is suspected, the patient would still need to be admitted to the hospital, for a range of non-invasive treatments. Even when catheterization or stenting is not an option, those with $K_{ij} = 1$ are still candidates for a range of blood-thinning medications (i.e., heparins) that give some more limited benefit for heart attack. At the very least, patients require close monitoring for arrythmias, because they would need immediate electrical shock if they developed. Because these patients will be admitted to the hospital even if they do not have heart attack diagnosed, Appendix 6.3, shows the results of an analysis where we exclude the 31.3% of patients who were admitted to the hospital with an ED diagnosis that indicates the emergency physician had not arrived at a diagnosis (i.e., those assigned primary diagnosis codes like 'chest pain'—symptoms, as opposed to a specific disease process). These patients, in whom a clear diagnosis had not been reached by the time the patient left the ED, might be suspected of heart attack (among other potential diagnoses). Of course not all these patients ultimately have heart attack. But by excluding them, we can calculate adverse event rates the remaining patients: those in whom the physician felt sure enough to assign an alternative diagnosis other than heart attack, as well as all those patients deemed safe to go home from the ED (and thus at very low risk of heart attack and other serious problems). This approach provides a lower bound for unsuspected adverse events. We find similar results in this population as our main sample, with a rate of adverse events 10.45% in the highest-risk bin, as opposed to 9.16% in the full population.

3.4.2 Patients Not Suspected of Heart Attack

Second, and building on this intuition, we isolate a group of patients in whom we know physicians are highly unlikely to have suspected heart attack, let alone suspected it highly and ruled out treatment: those who lack either an ECG or a troponin result in the ED. Because the ECG in particular is so inexpensive and non-invasive, it is done on everyone in whom the physician has a slight suspicion of heart attack.[39] Troponin, because it requires a blood draw, is somewhat more invasive but still very commonly done. Because both tests are also done for a range of other reasons unrelated to heart attack, their presence alone does not necessarily imply suspicion for heart attack—but their absence does imply lack of suspicion. Again here, even patients who are known to be treatment-ineligible would still have these tests done. If the physician suspects a patient is high-risk, but sees a contraindication to treatment, she will consider alternative, non-invasive therapies: there is still a known or potential blockage and that must be addressed. For these patients, the ECG and troponin are critical indicators of the risk of complications, and guide non-invasive treatments such as intravenous medication and admission to the hospital for observation.

In other words, if a patient lacks an ECG or a troponin, we know that $K_{ij} = 0$. The

[39]Indeed, it is often done by nurses at the triage desk, before the physician has even seen the patient.

specific assumption here is that, if the physician did not suspect a patient of heart attack enough to do an ECG or a troponin, she cannot have suspected that the patient was at extremely high risk for heart attack *and* further determined they were ineligible for treatment or testing with catheterization. In these patients, if our model predicts high risk, and we go on to see $A_{ij} = 1$, it is an unambiguous indication of true failures to test on the physician's part. If by contrast our findings were due to private information about treatment eligibility, patients without ECGs or troponin should have very low (base) rates of adverse events. The top two panels of Figure 2 first shows the fraction of patients lacking ECGs and troponin. These fractions decrease in predicted risk, but remain substantial in the highest-risk bins (which are now based on quintiles in the tested, rather than deciles, for sample size reasons): 26.5% lack an ECG (vs 80.7% in the lowest-risk bin), and 56.33% lack a troponin result (vs 95.1% in the lowest-risk bin). The bottom two panels show the adverse event rate in these high-risk patients, which is substantial. For patients in the highest-risk bin, the realized adverse event rate is 5.66% in those without an ECG, and 6.80% in those without a troponin. These rates are 1.01 percentage points lower and 0.14 percentage points higher than the 6.66% rate in the full population, respectively, and still well above the clinical threshold of 2% ($p = 0.024$ and $p < 0.001$ for those without ECGs and troponins, respectively). Together these results suggest that private information by physicians—either about risk of blockage or about suitability for invasive testing or treatment—cannot explain our findings.

3.5 Natural Experiment

The results so far provide some evidence for under-testing, but also leave some room for doubt. In Section 3.4 above, we ruled out the possibility that physicians were not testing the high-risk because they knew they were high risk but also knew they could not be treated. If that were the case, at a minimum they should have done an ECG and troponin test. But this does not show that these neglected high-risk patients would benefit from treatment. Now, there are no other known sources of treatment heterogeneity beyond life expectancy and suitability for treatment, both of which we have accounted for. Still, it would be nice to have more direct evidence that testing impacts health.

One strategy would be to compare long-term adverse event rates A_r^ℓ in tested vs. untested patients. The advantage of this is that, unlike our other health measures, it is formed similarly in tested and untested patients, over the period from day 30 to day 365 after the initial ED visit. A simple way to do this would be to form OLS estimates of how testing impacts these long-term health outcomes on average:

$$A_{ij}^\ell = \beta_0 + T_{ij}\beta_1 + \widehat{m}(X_{ij})\beta_2 + \mathsf{TimeControls}_j\beta_3 + \epsilon_{ij} \tag{3}$$

We might also wish to estimate the effect of testing separately, as a function of a patient's predicted risk

$$A_{ij}^\ell = \beta_0 + T_{ij}\beta_1 + \widehat{m}(X_{ij})\beta_2 + T_{ij} \times \widehat{m}(X_{ij})\beta_3 + \mathsf{TimeControls}_j\beta_4 + \epsilon_{ij} \tag{4}$$

Of course, the relationship between testing and health outcomes is confounded by selection.

So any health effects estimated via OLS could reflect the direct consequences of testing and treatment, or selection based on physician private information.[40]

To get around this selection, we would ideally like some kind of experiment, where we could actively test some untested high-risk patients at random, and measure the net effect on their overall health. While such an experiment is beyond the scope of the paper, there is natural variation in the data that might serve as a (limited) proxy for such an experiment.

3.5.1 Triage Shift Variation

The nursing staff assigned to triage patients when they arrive at the ED front desk rotates over the course of the day. As Chan and Gruber (2020) note, triage nurses have considerable discretion in many aspects of a patient's initial care, which can influence downstream decision making by physicians regarding testing. For example, a nurse can notice that a patient with chest pain is either 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 she brings the patient back into the room. As a result, we hypothesized that the testing rate (ultimately determined by the physician) could be affected by the particular make-up of the nursing team working the triage desk.

While we do not observe the identity of the nursing team, we do know the times at which a given shift begins and ends. So we calculate shift-level testing propensities via the following model:

$$T_{ij} = \beta_0 + \mathsf{Shift}_j\beta_1 + \widehat{m}(X_{ij})\beta_2 + \mathsf{TimeControls}_j\beta_3 + \epsilon_{ij} \tag{5}$$

Each visit's testing likelihood is modeled as a function of a vector of indicators indexing the particular triage shift she showed up in, modeled as a random effect. In addition, we control for a vector of time variables for visit j, that captures differences in testing rate attributable to the mix of patients showing up at a given time (i.e., fixed effects for year, week of year, day of week, and hour of day), as well as patients' predicted risk.[41] We end up with $5,925$ random effects that measure the testing propensity of each shift, and verify that the variance of these effects is non-zero ($p = 0.0492$) by running 1000 bootstrap simulations.

Figure 4 shows two checks of the validity of this setup. First, to test balance, we regress a set of pre-triage variables X_{ij} on the shift testing rate $\widehat{T_{\mathsf{Shift}j}}$, predicted from the random effect. (This is formed by regressing a testing indicator on the shift random effects, and predicting testing. This simply rescales the effects into units of testing probability, which we verify by also performing a regression of T_{ij} on $\widehat{T_{\mathsf{Shift}j}}$.)

$$X_{ij} = \beta_0 + \widehat{T_{\mathsf{Shift}j}}\beta_1 + \widehat{m}(X_{ij})\beta_2 + \mathsf{TimeControls}_j\beta_3 + \epsilon_{ij} \tag{6}$$

[40] Appendix 7 (Figure 6 and Table 11) shows these estimates. On average, and especially among lower-risk patients, rates of long-term adverse events are higher in tested patients. This would make sense if tested patients are tested by physicians because they are at unobservably higher-risk. When we move to the highest-risk patients, on the other hand, this trend reverses: the highest-risk untested patients have far higher rates of long-term adverse events than tested patients.

[41] The model is fit on the full sample (train and test cohort, using out-of-sample, cross-validated risk predictions for the train cohort) with the exclusions detailed above. A handful of observations for which we do not have a precise start time are dropped, so we have N = 238,459.

The top panel shows the results. As expected, the shift effects are linked to large variation in testing rate. We find some small, but statistically significant, differences in some of the pre-triage variables. For example, a 1 standard deviation (SD) increase in shift testing rate $\widehat{T_{\mathsf{Shift}}}$ implies 0.012 SD difference in predicted risk (p=0.0027). Considering differences in a set of indicator variables indexing bins of predicted risk, we can trace the source of this to a small but significant imbalance in the highest bin only (0.0103 SD, p=0.0103). However, the size of this effect is quite small: if we used the relationship between predicted risk and testing rate from the entire sample, this difference would account for only .01% of the total variation in testing across shifts. We found this reassuring regarding the extent to which risk differences are influencing shift testing effects. We also note small but statistically significant (or nearly so) differences in age (0.0095, p=0.0177), and the likelihood of self-reporting Hispanic ethnicity (0.0070, p=0.0814). The bottom panel shows that, net of all controls, there is wide residual variation in the mean predicted risk of patients triaged in a given shift, at any level of the shift's testing rate.

Beyond looking at pre-triage observables, another way to test whether patients differ meaningfully in their risk across shifts is to consider a 'downstream' measure: for those tested at a given level of predicted risk, what is the realized yield? This approach might detect differences in risk driving differences in testing yield, indicating imbalance on unobservable factors. For example, if higher testing rates in some shifts reflected true but unobserved differences in patient risk, we would expect to see higher yield. This would cast doubt on the assumption of as-if-random variation across shifts, because higher testing rates would reflect true risk differences that are simply not captured in our predictions. Table 4 shows the results of comparisons of yield across quartile bins of shifts, based on estimated testing random effects. Despite wide variation in testing rate (7% in the top vs. 1% in the bottom quartile), there is no clear correlation between the testing rate of shift, and the yield of testing in that shift. While standard errors are large, and we cannot definitively rule out correlations, this argues against a large influence of unobservable patient risk factors driving testing rates.

3.5.2 Are Physicians Failing to Test Patients Who Benefit From Testing?

After verifying that variation in testing rates is plausibly unrelated to patients' risk, we turn to estimating the effect of assignment to testing on health outcomes. We begin by looking at the average effect of testing, across all patients, which we estimate by comparing patients arriving to the ED on a higher- vs. lower-testing triage shift. Given sample size considerations, we use the full sample here (N = 213,484) not just the hold-out (results in the hold-out are very similar, just less precise as we would expect). To generate out-of-sample risk predictions for all observations, we perform five-fold cross-validation, training the model on 80% of the data and applying it to form risk predictions on the held-out 20%, then repeat 4 times. Analogous to the OLS formulation in (3) we estimate:

$$A_{ij}^{\ell} = \beta_0 + Q(\widehat{T_{\mathsf{Shift}j}})\beta_1 + \widehat{m}(X_{ij})\beta_2 + \mathsf{TimeControls}_j\beta_3 + \epsilon_{ij} \tag{7}$$

That is, we look at the effect of assignment to higher or lower testing rates on health outcomes. Our measure of assignment to testing is a vector of indicators for each quartile of the triage shift random effect for testing. $Q(\widehat{T_{\mathsf{Shift}j}})$, estimated in Equation (5). The range from lowest to highest quartile of testing rates is 0.5% to 6.6%. We also test a linear specification to ensure that our results are not driven by a specific functional form. We hold constant the same vector of time controls used to estimate the shift effects, and the patient's predicted risk.

Table 5 shows the results: largely null. There is no average effect of increasing testing on health, whether measured by total adverse events from day 31-365 after visits (Column 1), or on either sub-component of the outcome: diagnosed heart attack or arrythmia (Column 2) or death (Column 3). These results re-create the well-known result in the literature: "flat of the curve" medicine, where increases in testing yield little benefit (Fisher et al., 2003). We can estimate of cost-effectiveness of increasing testing in this way, by simply re-estimating Equation 7 with life years and costs as the dependent variables:

$$\eta_{ij} = \beta_0 + \beta_1 Q(\widehat{T_{\mathsf{Shift}j}}) + \beta_2 \widehat{m}(X_{ij}) + \beta_3 \mathsf{TimeControls}_j + \epsilon_{ij} \tag{8}$$

$$c_{ij} = \gamma_0 + \gamma_1 Q(\widehat{T_{\mathsf{Shift}j}}) + \gamma_2 \widehat{m}(X_{ij}) + \gamma_3 \mathsf{TimeControls}_j + \epsilon_{ij} \tag{9}$$

Here, as above, η_{ij} reflects remaining life-years, which are set to zero if the patient dies, and adjusted by the usual discount rate if the patient has a diagnosed heart attack; and $c_{ij} = c_{Tij} + S_{ij}c_{Sij}$, the total costs of testing and treatment in the 10 days after the index visit. Once the marginal life-years saved (or lost) and costs are estimated, we can calculate the marginal cost-effectiveness of the additional tests simply by dividing γ_1/β_1. These results are shown in Appendix 8. As we might expect given the very small differences in long-term health outcomes, increases in testing actually offer negative benefit on average. That is, moving the average patient from a low- to high-testing regimen will cost $894, but the patient's life expectancy will actually decrease by 1.46% (or, 0.551 years).

The presence of under-testing as identified by the algorithm, however, suggests we should see benefit for subsets of patients. In particular it suggests that, were we to look at the impact of testing by risk, we ought to find health effects. To capture this, we re-estimate (7), but this time allowing the effect of testing on long-term outcomes to vary by a patient's predicted risk, analogous to (4) above:

$$A^{\ell}_{ij} = \beta_0 + Q(\widehat{T_{\mathsf{Shift}j}})\beta_1 + \widehat{m}(X_{ij})\beta_2 + Q(\widehat{T_{\mathsf{Shift}j}}) \times \widehat{m}(X_{ij})\beta_3 + \mathsf{TimeControls}_j\beta_4 + \epsilon_{ij} \tag{10}$$

Table 6 shows the results. Column (1) shows the effect of arriving during a high- vs. low-testing shift. Across all specifications of the model (linear testing rate and risk quintiles; testing rate quartile and linear risk; or testing rate quartile and risk quintiles, in Appendix 8), we find large reductions in adverse events in high-risk patients, and none in low-risk patients. Of course, because there are far more low-risk patients (recall that the risk bins reflect quantiles formed in the tested), the average effect is null, as in Table 5. But thanks to model predictions, we can compare, for example, high-risk patients arriving during a top testing quartile shift (average testing rate: 6.6%), relative to a bottom testing quartile shift

(average testing rate: 0.5%). For those whose predicted risk would put them in the top quintile of the tested, we find a decrease in adverse events of 4.4 percentage points—a 35.8% reduction relative to patients in the bottom risk quintile, who have no decrease in event rates. We find similar results for each component of adverse events: high-risk patients suffer 3.41 percentage points fewer diagnosed heart attacks (a 53.3% decrease), and 1.21 percentage points fewer deaths (a 17.1% decrease). Low-risk patients again have no change in these event rates. (These numbers are from the fully non-linear specification in Appendix Table 12.) By translating these numbers into cost-benefit calculations as above, we can also show that testing high-risk patients is highly (marginally) cost-effective at $53,211 per life-year—a striking contrast to the very low cost-effectiveness we would have concluded had we failed to segment the population by predicted risk.

One possible concern with these numbers is that the adverse events we study are all measured over the 30-365 days after ED visits, excluding the first month. We do so because tested patients are mechanically more likely to be diagnosed with heart attack than untested patients, simply by virtue of being in the hospital for testing. To ensure that these results are not simply being driven by the choice of the interval, we rely on the fact that death, thanks to our linkage to Social Security data that captures death outside of the hospital, does not suffer from this difference in ascertainment. So we re-run the analysis with death as the outcome, measured over the entire year after visits including the first month, in Column (4) of Table 6. We find a similar 1.46 percentage points fewer deaths (a 13.8% decrease) in high-risk patients, and no change in low-risk patients.[42]

These results show that testing predicted high-risk patients does yield high returns. In other words, physician private information about treatment heterogeneity cannot account for our findings. Of course, this does not imply that *all* high-risk untested patients would benefit the same: we are constrained by the extent of variation in testing rates in our data. This experiment does not allow us to say anything about those patients that would never be tested even in the highest testing shift-rate. Still, combined with previous findings, these results establish that alongside over-testing, physicians also under-test, failing to test some patients the algorithm would have correctly identified as high risk.

4 Why do Physicians Make Testing Errors?

So far, we have documented the welfare costs of physician errors, but not answered an important question: why do physicians make these errors? To better understand why physicians mis-predict, we build a second predictive model: for each patient, we predict whether the physician will choose to test them that patient. This algorithmic model of physician choice can then be contrasted, patient by patient, with the algorithmic model of actual risk. Thanks to the analyses above, which validates model predictions with respect to real health outcomes, we can provide suggestive evidence on decision-making errors: patients for whom

[42] Appendix Table 12 shows similar results where testing rate is specified as a set of indicators for shift quartiles instead of a linear term.

testing deviates not from true risk. Our approach builds on a long tradition of research 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, Ferris, Tu, Hux, and Schull, 2001).

We investigate two kinds of error commonly discussed in the decision-making literature. First, one set of models emphasizes 'boundedness': physicians may not be able to attend to, process, or mentally represent the rich set of data available on their patients. As a result, they may resort to a simpler model of risk than is statistically optimal. By contrast, a second set of models emphasizes errors: even if all variables were considered, physicians may make statistical reasoning errors and give undue weight to some over others.

4.1 Boundedness in Physician Judgments

We develop an empirical test that builds on the observation that boundedness is analogous to the concept of regularization in machine learning (Camerer, 2018). In forming predictions, regularization is used to control overfit. It does so by parameterizing the complexity of models. In LASSO, for example, the number of variables with non-zero coefficients (or more precisely the sum of the absolute size of coefficients) is a measure of complexity. To arrive at their final predictors, algorithms first fit a different predictor for each level of complexity; and then they choose a level of complexity based on which results in the best out-of-sample fit. These intermediate models, though, provide a way to test for bounded rationality: they provide a sequence of models of actual risk that vary in complexity.[43] These allow us to ask the question: which level of model complexity best predicts physician testing decisions?

We begin with the LASSO model of risk we developed as one component of our full ensemble model, because its complexity measure is very transparent. The process of regularization for the LASSO consists of jointly minimizing two quantities: the residual sum of squares, and the sum (specifically, the L1-norm) of the model coefficients. The latter sum has a weight in the objective function that is allowed to vary, and the typical process of tuning the model selects the optimal weight, between zero (where the objective function is simply OLS) and infinity (where all model coefficients are zero) via cross-validation. But rather than simply retaining the single best model that minimizes the objective function, we preserve all models along the regularization path. This provides for $k \in [0, 1500]$, the best fit linear model that uses at most k non-zero coefficients.[44]

We then correlate physician testing decisions with each of these risk models individually. Doing so allows us to calculate the complexity of the LASSO risk model that explains the

[43] We will use 'risk' to denote our usual quantity of yield of testing in the tested, because all our results so far indicate that this is highly correlated with risk. However, despite the generally reassuring results presented above indicating that this correlates with true risk, all the caveats regarding unobserved risk factors apply.

[44] Given the computational expense of extending this to the full set of $k = 16,381$, we verified that results continue as expected at specific points, but do not show results for the full path. We emphasize that, even if the particular set of variables the LASSO chooses is somewhat arbitrary in the setting of correlated, noisy input variables, the overall complexity of these models is a more stable quantity that does not depend on the particular variables chosen (Mullainathan and Spiess, 2017).

most variance in physician decisions, which we denote k_h^*. We can of course also calculate the complexity of the LASSO risk model that best predicts risk, which we denote k_r^*. Boundedness implies that $k_h^* < k_r^*$: physicians rely on a simpler model of risk than is optimal. This exercise, naturally, quantifies the complexity of physicians' risk predictions on the basis of observable factors only. Physicians, as we have emphasized, also have access to information that is unobserved in our data; and we have no way of investigating whether these are being neglected. That said, we can test hypotheses related to boundedness conditional on the large set of variables we do observe.

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 10.5 shows similar results with 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 the same but for predicting physician decisions. For risk, we see that R^2 increases at first then decreases as additional variables lead to over-fitting. For physician decision, 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 risk, the empirical optimum is at 224 variables, while for physician decision making it is at 49 variables.

To statistically test the implication of this figure, we estimate two regressions. First we define the risk predictor that uses only 49 variables as $\widehat{m}_{\mathsf{simple}}(X_{ij})$ and the more complex one as $\widehat{m}_{\mathsf{complex}}(X_{ij})$. We can decompose

$$\underbrace{\widehat{m}_{\mathsf{complex}}(X_{ij})}_{\text{Best for Risk: } k_r^*} = \underbrace{\widehat{m}_{\mathsf{simple}}(X_{ij})}_{\text{Physician: } k_h^*} + \underbrace{(\widehat{m}_{\mathsf{complex}}(X_{ij}) - \widehat{m}_{\mathsf{simple}}(X_{ij}))}_{\text{Incremental}}$$

We then regress both the testing decision (in all patients) and the realized yield (in the tested) on both the simple risk model and the incremental risk prediction of the complex model:

$$T_{ij} = \beta_0 + \beta_1 \widehat{m}_{\mathsf{simple}}(X_{ij}) + \beta_2[\widehat{m}_{\mathsf{complex}}(X_{ij}) - \widehat{m}_{\mathsf{simple}}(X_{ij})] + \epsilon_{ij} \tag{11}$$

$$S_{ij} = \gamma_0 + \gamma_1 \widehat{m}_{\mathsf{simple}}(X_{ij}) + \gamma_2[\widehat{m}_{\mathsf{complex}}(X_{ij}) - \widehat{m}_{\mathsf{simple}}(X_{ij})] + \epsilon_{ij} \tag{12}$$

This regression allows us to statistically test boundedness. Under this view of physician decision-making, we would hypothesize that incremental risk does not predict testing, and thus that we cannot reject $\beta_2 = 0$. In addition, we would also hypothesize that incremental risk has contains enough true signal that we can reject $\gamma_2 = 0$. These results are shown in Table 7. 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 marginal contribution to predicted risk. Column (2) shows that conditional on the simple model, the additional risk information in the complex model is not predictive of testing - the coefficient is both very small and statistically insignificant. In Column (4) we repeat this exercise but with yield, where we see that the additional risk information is highly significant. These results provide suggestive evidence in support for boundendess: more complex information that predicts risk is being omitted from the physician's testing decision.

This phenomenon is not specific to the hospital setting we study: we see similar results in our Medicare data, where $k_r^* = 299$ variables, while $k_h^* = 21$ (Appendix 10.6). Nor is it specific to a linear model, or the LASSO's particular regularization function. We repeat this exercise, with some modifications, for the gradient boosted tree model that forms the other part of our ensemble. Specifically, we measure model complexity using the number of iterations of the model: each successive model is fit to the residual of the last, progressively increasing the complexity of the model. We show the results of this exercise in Appendix 10.5: the R^2 of increasingly complex models is flat and then decreasing for predicting the physician's testing decision after the 7th iteration, while the best model for predicting blockage is at the 12th iteration. While we are unable to perform the formal test we devised for LASSO (because individual trees have considerable randomness, we calculate R^2 at a given iteration for an average of 100 separate tree models) we can compare mean performance across these 100 models. We find that the R^2 increases from iteration 7 to iteration 12 for blockage (0.170 to 0.174, t-test: p = 0.00011), but not for testing decisions (0.0390 to 0.0384, p = 0.740).

This illustrates the 'bounded' part of 'bounded rationality,' but does not speak to the 'rationality' part. The presumption in boundedness models is that, given the information a decision-maker does use, they use it correctly. To examine this, we focus on the variables chosen by LASSO for the simple model above $\widehat{m}_{\mathsf{simple}}(X_{ij})$. For each variable, we correlate it with test outcome and then correlate it with testing decision. Figure 6 shows the scatterplot for all 49 variables of these two correlations: correlation with the testing decision on the y-axis, and its correlation to blockages in the tested on the x-axis.[45] We find that the variables are largely weighted proportional to their observed relationship to true risk: the relationship appears linear and positive, and the R^2 is 0.433. In other words, physicians largely get the signs and magnitudes of these variables right. This is clearly only suggestive and is far from a precise quantification of rationality: much more formal econometrics is needed to produce a formal test along these lines. But as a descriptive statistic, it provides some suggestive evidence.

Together, these results provide some support for boundedness and rationality in physician judgments. Physicians, much like algorithms, appear to be regularizing (Camerer, 2018; Gabaix, 2014): they identify a small number of good risk predictors and use them if not perfectly, at least quite well. But conversely, they appear to neglect hundreds of other variables that, while individually small, together account for much of the true risk model's explanatory power. It is worth noting that our findings generally agree with a long tradition of research since Dawes, Faust, and Meehl (1989), finding that actuarial models can outperform clinical judgment. However, a notable difference relates to model complexity. Dawes, Faust, and Meehl (1989) raises the possibility that clinicians use too complex a model; and that a simple statistical model would do better. Our story is different: it is physicians who are using a model that is too simple; and a statistical model does better by being more complex.

[45] We standardize the test, yield, and predictor variables, and run test and yield on the predictors in a set of univariate regressions, so that each regression coefficient gives us the correlation coefficient and its standard errors.

4.2 Biases in Physician Judgments

We next turn to systematic biases. We already see one suggestive piece of evidence that biases might exist alongside generally rational weighting of variables, in Figure 6: one point (Reason for visit: chest pain), clearly lies further up than all the others. When a patient comes to the ER complaining of chest pain, that fact is informative of heart attack. Yet it appears to correlate far more with test than is merited by this information, at least in this figure. Put differently, this data point raises the possibility that physicians over-weight chest-pain: placing more weight on it in deciding who to test that its informativeness for merits. We can test this in a straightforward way by regressing testing on predicted risk as well as chest pain. We would be asking whether chest pain predicts testing above and beyond any correlation with actual risk:

$$T_{ij} = \beta_0 + \beta_1 \widehat{m}(X_{ij}) + \beta_2 \mathsf{ChestPain} + \epsilon_{ij} \tag{13}$$

$$S_i = \gamma_0 + \gamma_1 \widehat{m}(X_{ij}) + \gamma_2 \mathsf{ChestPain} + \epsilon_{ij} \tag{14}$$

This exercise shows that in fact a patient with chest pain is 14.2 percentage points more likely to be tested, even conditional on predicted risk. Of course chest pain is a strong signal - those with chest pain are 61.12% (6.8 percentage points) more likely to have acute blockages - but physicians appear to over-weight this strong signal. This result generalizes to a broader range of symptoms, beyond chest pain, as shown in Appendix 10.8. Of the 10 most common symptoms in our data, nine are significant predictors of the testing decision, including chest pain and shortness of breath (large and positive), and several other smaller negative predictors (e.g., abdominal pain).[46]

4.2.1 Symptom Salience

This result raises the question of whether symptoms as a category are over-weighted. These are the most immediate thing the physician sees and thus salience could be a mechanism of errors in judgment (Tversky and Kahneman, 1981; Bordalo, Gennaioli, and Shleifer, 2012). We test for this by creating a new risk predictor, $\widehat{m}_{\mathsf{symptom}}(X^s_{ij})$, which uses as its inputs solely the vector $X^s \subset X$ of patient symptoms (as indicators). This prediction is formed exactly as our usual $\widehat{m}(X_{ij})$ in the training set except the model only uses data on symptoms. In the holdout set, we then regress the test decision on both the full risk predictor and the symptom-based risk predictor together; and as usual, verify that the symptom-based risk predictor is not additionally predictive of yield).[47]

$$T_{ij} = \beta_0 + \beta_1 \widehat{m}(X_{ij}) + \beta_2 \widehat{m}_{\mathsf{symptom}}(X^s_{ij}) + \epsilon_{ij} \tag{15}$$

$$S_{ij} = \gamma_0 + \gamma_1 \widehat{m}(X_{ij}) + \gamma_2 \widehat{m}_{\mathsf{symptom}}(X^s_{ij}) + \epsilon_{ij} \tag{16}$$

[46] Abaluck, Agha, Kabrhel, Raja, and Venkatesh (2016), while they lacked data on symptoms present at the visit itself, investigated whether patients with previous symptoms were over-tested, and found evidence consistent with a similar bias.

[47] For this and subsequent regressions, we control for a vector of risk bins, as well as the linear risk, to account for non-linearity of risk in predicted risk. We show the coefficient on the linear term but omit the others for simplicity.

Results from Equation 15, exploring testing, are in Table 8. Column (1) shows a simple regression of testing on predicted risk, for reference. Column (2) then adds the symptom-based predictor. This shows that physicians put considerable weight on risk information found in symptoms, even after conditioning on the full risk predictor. While this suggests that symptoms as a category appear to be highly over-weighted, we also test this hypothesis more formally, relative to other categories of risk information available to the physician. To do so, we form additional risk predictors, similar to the symptom-based predictor, using the various other types of data we have available to us: demographics, structures diagnosis and procedure codes, laboratory data, vital signs, demographics, and medication. We then add these as predictors in the Equation 15, and show the results in Column (3). Risk information from symptoms is indeed a major predictor of the physician testing decision relative to other categories. Demographics and prior diagnoses are also weighted highly, while risk information from prior laboratory studies, vital signs, and medications are actually negatively correlated with the testing decision. Appendix 10.3 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, while self-reported Hispanic patients are under-tested. In Appendix 10.7 (Table 18), we show results from Equation 16 the yield of testing on these risk predictors, and find that none are additionally useful for predicting yield (as expected, given that the full predictor also can use symptoms). Of course, this test is not definitive: there could be a complex relationship of symptoms with unobservable risk factors that would make this weighting not clearly evidence of error. But it does suggest models of systematic errors in human judgment that deviate from the rules of probability (Tversky and Kahneman, 1974), alongside our earlier results consistent with boundedness.

4.2.2 Symptom Representativeness and Stereotypes

Salience is just one possible mechanism by which symptoms can be over-weighted. Certainly, symptoms as a category are plausibly the most salient. Yet there may be more structure to these symptoms - not all symptoms are diagnostic of heart attack. One such explanation we explore is related to the idea of representativeness (Tversky and Kahneman, 1974), as formalized in the model of stereotyping of Bordalo, Coffman, Gennaioli, and Shleifer (2016). This model predicts that a physician, tasked with forming probability judgments for a patient with symptom M, will (*ceteris paribus*) not test the symptom's objective likelihood of coronary blockage $Pr(B = 1|M = 1)$. Rather she will judge according to $Pr(B = 1|M = 1) \times h\big(\frac{Pr(M=1|B=1)}{Pr(M=1|B=0)}\big)$ where $h(\cdot)$ is a monotone function. So holding constant true risk of blockage, the judged probability also depends on the distinctiveness of a symptom for heart attack relative to other possible causes of that symptom.

This model has a crisp empirical prediction: at the same predicted risk, patients with more or less representative symptoms are more or less likely to be tested.[48] We begin to

[48]This possibility is related to several observations from the clinical literature on diagnostic error (reviewed in IOM (2015)), most notably the idea of "premature closure" (Croskerry, 2002; Graber, Franklin, and Gordon, 2005). When physicians seize on one condition that might explain a patient's symptoms, they may be tempted to "close early" and fail to consider another.

investigate this by creating a list of symptoms that are potentially representative of heart attack: in those tested patients ultimately found to have heart attack, we look back at the presenting symptom. We limit to those with appreciable frequency, in this case over 0.5%, and show the 16 symptoms meeting this criterion in Appendix 10.6 (Table 17). Along with the overall frequency of these symptoms, we also calculate the representativeness of each symptom for heart attack: $\frac{Pr(M=1|B=1)}{Pr(M=1|B=0)}$. This identifies 9 symptoms with a ratio over 1, which we consider representative of heart attack: 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 heart attack, or because they were found unresponsive or in cardiac arrest by paramedics). The remaining 7 symptoms are more common in the general population than in those with heart attack (e.g., dizziness, nausea).

With this framework in place, we then form another version of symptom-based predicted risk $\widehat{m}_{\mathsf{symptom}}(X_{ij}^s)$—but this time formed solely from the subset of 9 *representative* symptoms. Column (4) of Table 8 shows the results of adding this to the regression we described previously (Column 3) 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). Appendix Table 18 confirms this new predictor, like the others, has no incremental value for predicting yield of testing. 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 heart attack.

5 Implications of Physician Error for Health Policy

Policy makers have long viewed health care through the lens of misaligned incentives: doctors and health systems are paid based on quantity of care provided, so they seek to provide more care. This leads to a view where over-use—low-value care that provides little benefit for its cost—is the primary source of inefficiencies. It also leads to a set of policies aimed at correcting these misaligned incentives. Incentives alone, though, cannot explain why doctors would fail to test high-risk patients but also test low-risk ones. Our results contribute to a growing body of work that argues that an incentive-focused perspective on health care is incomplete; and puts forward richer models of physician behavior.

5.1 Richer Models of Physician Behavior

Two recent papers are worth mentioning in particular. Chan, Gentzkow, and Yu (2019) highlight the importance of skill differences: some physicians are better than others in deciding whom to test. Chandra and Staiger (2020) highlight comparative advantage: some health systems specialize and focus on certain tests and conditions. These new channels have implications for interpreting the patterns of testing we see. What appears to be over-testing, for example, may not be driven by physicians' incentives to test. Instead, lower-skill physicians

who are less able to distinguish heart attack from benign causes, may rationally test more to compensate. Alternatively, hospitals that see more heart attacks, and less pulmonary or oncological disease, may test more efficiently (though we cannot test this in our data). All of these perspectives caution us from interpreting cross-sectional variation as evidence of over-use driven by the private benefit physicians receive for providing more care.

These models all have a key implication for policies aimed at making health care more efficient: incentivizing physicians to change their behavior could be highly counter-productive. The apparent inefficiencies we see are in fact manifestations of physicians optimizing on some important margin, that is currently not obvious to policy makers. In Chandra and Staiger (2020), decisions that look suboptimal are in fact efficient because of hospitals' comparative advantage. In Chan, Gentzkow, and Yu (2019), lower-skill physicians, who are less able to distinguish serious from benign cases, rationally test more to compensate. So policy makers hoping to get physicians to change their behavior could end up forcing hospitals specializing in one kind of care to provide another; or forcing lower-skilled physicians to reduce testing without knowing how.

Because our risk predictions are formed ex ante and at the individual patient level, they offer a natural way to explore the relationship of physician skill and testing rate. Specifically, we model the testing decision as a function of patient predicted risk, as well as a set of effects for the skill and testing rate of physician d, for the 70 physicians in our sample with more than 500 recorded patient encounters:

$$T_{ij} = \alpha_d + \theta_d \widehat{m}(X_{ij}) + \epsilon_{ij} \tag{17}$$

We assume skill is captured by the physician-specific slope vector θ_d—how much does risk predict testing—and general testing propensity by the physician-specific intercept vector α_d. Figure 7 then plots these two measures against each other, where each point indicates a different physician. Testing propensity α_d is on the y-axis and skill θ_d is on the x-axis; and the confidence intervals for each of these is represented by the ovals. As is clear in the figure, there is a strong negative correlation between these two physician characteristics. These results suggest as proposed by Chan and Gruber (2020) that high-testing physicians are of lower "skill", at least as measured by correlation with algorithmic risk predictions.[49]

5.2 Physician Error And Low-Value Health Care

Our results provide a complementary, but distinct, behavioral account of health care inefficiencies: physician error, due to boundedness and bias in risk prediction. In this sense, they resonate most with the approach taken by Abaluck, Agha, Kabrhel, Raja, and Venkatesh (2016), who also find both over- and under-testing for pulmonary embolism. Like the models of Chan and Gruber (2020) and Chandra and Staiger (2020), the channel we highlight has little to do with incentives (although it certainly allows for the fact that incentives can push

[49] Appendix 10.4 performs a similar exercise, but this time focusing on how many total patients a physician sees. There we find suggestive evidence that higher volume of patients leads to better decisions– more specifically, a larger correlation of testing with risk predictions.

physicians to test more). But in contrast to these two models, it does not focus on physicians optimizing on some other margin. Rather, we emphasize error: large welfare losses arise because of mis-prediction of patient risk, rather than as an unintended consequence of physicians optimizing on some other margin.

Our results, too, have important implications for how to address low-value health care, and in particular a common policy prescription in health: getting high-utilizing hospitals to act more like low-utilizing ones. Many current policy initiatives try to incentivize physicians to reduce care, by changing prices or reimbursement schemes. These efforts are grounded in traditional moral hazard models that assume—without empirical evidence—that physicians will reduce low-value tests when they are incentivized to test less. But our results suggest that physicians will only reduce tests that they *perceive* as low-value. So much as the existence of patient error changes optimal insurance design in 'behavioral hazard' models (Baicker, Mullainathan, and Schwartzstein, 2015), allowing for the possibility of physician error changes the calculus for structuring interventions on physician behavior.

5.2.1 Risk Distribution of Physicians' Marginal Patients

Our data gives a glimpse of how incentivizing physicians to cut back on testing can have perverse consequences. Implicit in the results of our natural experiment, above, is the idea that physicians mis-rank patients by risk and in so doing leave large welfare gains on the table. As we move between high- and low-testing regimes, the additional tests are on average not of high value (columns 1 to 3 of Table 6). And yet, at the same time, there are also high-risk patients that remain untested when testing rates increase, even though they would benefit. This setup allows us to ask: which patients do physicians themselves view as 'marginal'—when they decide to test more, whom would they test? Our empirical strategy here is simple: we inspect specific cells of predicted risk, and quantify the extent of variation in testing in a given risk cell. If physicians rank-order patients more or less correctly by predicted risk, we should find that marginal patients fall into a narrow band of risk: the range where the private benefit of testing to the physician causes them to test less risky patients than is optimal. For example, consider the most conservative shifts, those in which the fewest patients are tested. The average rate of testing is only 0.5%; among the highest-risk patients (top quintile), 15.2% are tested, while in the rest of the population only 0.35% are tested. Now imagine if physicians increased overall testing, to the rate of the next highest-testing quintile: 1.40%. Under moral hazard, this increase should be concentrated among lower-risk patients than are currently being tested. So we would expect to see increases in testing in the next-highest risk group, just lower than those patients being tested in the most conservative shifts. We certainly would not expect to see increases in testing in the highest-risk group: even in the lowest-testing regimes, high-risk patients who need to be tested are already being tested. By contrast, under a more behavioral model, we could see increases in testing anywhere in the risk distribution where physicians mis-predict risk.

The top panel of 8 shows the probability of testing in a given cell of predicted risk, across the four quartiles of testing rate based on triage shifts, in our hospital sample. We find that, when physicians test more, they test *everyone* more. Increases in testing happen

in all cells of predicted risk, implying that marginal patients are drawn from across the entire risk distribution, not just progressively lower-risk groups. To make our results more directly comparable to a large literature on geographic variations in care, we replicate this analysis using Medicare data. We first calculate overall testing rate using our Medicare sample for all US hospitals, then we sort hospitals into quintiles based on their testing rate. We then calculate the probability of testing in each bin of predicted risk, and replicate the graph where we can consider how testing rate changes in each risk bin. This shows the same result: hospitals that test more test everyone more. This exercise is inspired by the literature, where it is common in to make cross-sectional comparisons in care across geographic units, but naturally these comparisons can be confounded by unobservables. So in the Appendix 9.3, we identify a 'natural experiment' similar in spirit to our shift-based testing perturbations. We find that patients coming to the ED on a Thursday vs. Friday, or a Sunday vs. Saturday, are more likely to be tested. This is because tests are often performed after one night of observation for stability, and many hospitals do not routinely staff testing facilities on weekends, constraining the availability of tests. We find similar results in this setup (Appendix Figure 9b). Overall, whether we look in depth at a particular hospital or across all hospitals in the county, we find the same empirical fact: marginal patients are as likely to be high- or low-risk. As a result, incentivizing physicians to test less will cause them to throw the baby out with the bathwater, and sacrifice high-value care along with low-value care.

5.2.2 Comparing Over- and Under-Testing

The evidence so far demonstrates that physicians simultaneously over- and under-test. A natural question is, how do these two groups compare in size? To make this question precise, we consider two kinds of counterfactual policies. First, to identify over-testing we look for welfare-improving policy counterfactuals that use ex ante characteristics to identify tests that could be cut. The number of tests cut by such a policy produces one estimate of "over-testing". This exercise is straightforward. Using commonly accepted cost-effectiveness thresholds, we would cut 49.1% of all tests currently done by physicians, suggesting over-testing is quantitatively large.

Second, and in a parallel approach, to identify undertesting, we look for welfare-improving policy counterfactuals that use ex ante characteristics to identify untested patients who would benefit from testing. The number of people untested identified by such a policy produces one estimate of "under-testing". The challenge here, as we have noted, is in validating that a counter-factual policy would in fact be welfare-improving. To address this, we build on the two approaches we have already taken to this problem. In Figure 2 we used the adverse event rates as a benchmark for who should be tested. We see from that Figure that many of the highest-risk bins of untested patients have an adverse event rate in excess of the clinical threshold: a full 14% have a rate significantly above 2%. This gives us one estimate of under-testing. But a more conservative approach here would be to focus only on those patients without an ECG, as we did in Figure 3, which sets a lower-bound of sorts. This approach indicates that the top 2.68% of untested patients without an ECG (or, 1.93% of all

untested patients) have an adverse event rate above 2%. Testing all of these patients would increase testing by 63.1% (relative to the current rate). Our second approach to quantifying under-testing uses the results of the natural experiment, which reveal the health impacts of testing. By directly estimating the impact of testing assignment on costs and life-years, we can identify those predicted risk cells where testing would be marginally cost-effective (at $150,000 per life-year). For all patients in these bins, we then set the counterfactual testing rate to the maximum testing rate observed in the data (i.e., the highest-testing quartile of triage shifts). This policy would lead to increases in testing for a total of 1.16% of the untested, equivalent to increasing testing 37.3% relative to the current rate. Because this counterfactual is grounded in the effect of testing on health outcomes, and incorporates a concrete cost-effectiveness threshold, this is our preferred specification for estimating under-testing. Putting this together with our estimate of over-testing above (49.1% of current tests), our counterfactual policy would cut testing on net by 11.8%—but of all the tests recommended under this policy, 42.3% would be high-value *new* tests, done for high-risk patients physicians are not currently testing.

6 Conclusions

Much of our understanding of the health care system has its roots in how we model physician behavior. The fact that over- and under-testing coexist in our results speaks to the existence of errors in judgment on the part of physicians as well. Of note, this happens despite training and motivation to make use of extensive data noted in other research (Kolstad, 2013). As such it mirrors increasing evidence of widespread inefficiencies observed in patients' decision making, that that our current models cannot explain (Baicker, Mullainathan, and Schwartzstein, 2015; Brot-Goldberg, Chandra, Handel, and Kolstad, 2015; Handel and Kolstad, 2015).

This has implications for how interventions can be devised to reduce low-value care. To date, policy makers have aimed squarely at reducing overuse, grounded in rudimentary theories of moral hazard. This has taken the form of exhorting high utilizing doctors or hospitals to be more like their low utilizing neighbors (e.g., Liao, Fleisher, and Navathe (2016)), or by changing provider incentives (e.g., Obamacare) to reduce the volume of care delivered – and let doctors decide where and when to cut back care (e.g., Loewenstein, Volpp, and Asch (2012)).[50] These interventions have produced, at least on the patient side, a mixed record of targeting low-value care. More often, as the seminal RAND health insurance experiment and more recent work since has shown, changing incentives cuts all care – not just low-value care (Newhouse and Group, 1993; Brot-Goldberg, Chandra, Handel, and Kolstad, 2015; Chandra, Gruber, and McKnight, 2010). Our results on mis-prediction strongly suggest that similar problems affect interventions aimed at physicians, meaning these policies may have less impact than hoped.

[50]More nuanced views from economics draw attention to other mechanisms of over-use, stemming from patient selection into treatment, for example, Einav, Finkelstein, Oostrom, Ostriker, and Williams (2019).

While new research in economics has offered compelling new explanations for apparent over- and under-use besides incentives (Chandra and Staiger, 2020; Chan, Gentzkow, and Yu, 2019), these explanations invoke deeply rooted aspects of hospital practice, or physician quality. These factors may be hard to change—and it is not obvious that changing them is the right thing, since they result from hospitals and physicians optimizing on different margins (e.g., comparative advantage in a certain procedure, or control of false-negative rates). Mis-prediction, by contrast, is an error; it most closely echoes research invoking errors in physician judgment, whether in economics or medicine (Abaluck, Agha, Kabrhel, Raja, and Venkatesh, 2016; Graber ML, Wachter RM, and Cassel CK, 2012). The fact that we use electronic health data, which are readily available in all hospitals at the time of decision making, also opens up new channels for targeted interventions in clinical contexts, which could nudge providers to make better decisions. Interventions that improve the practice of medicine, rather than ones that simply change the incentives to practice it in a certain way, could be a powerful policy lever to drive efficient health care use.

The ability to form accurate, tailored risk predictions was a key part of building this evidence. This illustrates that machine learning has an interesting role to play both in applied decision making, and in testing theories in social science (Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan, 2017): comparing idealized predictions to the actions of individual decision-makers is a fascinating new lens through which to view human behavior in complex environments.

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.

Abelson, Reed and Julie Creswell (2012). "Hospital chain inquiry cited unnecessary cardiac work". In: *New York Times* 6.

Acemoglu, Daron and Amy Finkelstein (2008). "Input and technology choices in regulated industries: Evidence from the health care sector". In: *Journal of Political Economy* 116.5. Publisher: The University of Chicago Press, pp. 837–880.

Al-Lamee, Rasha, David Thompson, Hakim-Moulay Dehbi, Sayan Sen, Kare Tang, John Davies, Thomas Keeble, Michael Mielewczik, Raffi Kaprielian, and Iqbal S. Malik (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.

Antman, Elliott M., Marc Cohen, Peter JLM Bernink, Carolyn H. McCabe, Thomas Horacek, Gary Papuchis, Branco Mautner, Ramon Corbalan, David Radley, and Eugene Braunwald (2000). "The TIMI risk score for unstable angina/non–ST elevation MI: a method for prognostication and therapeutic decision making". In: *Jama* 284.7, pp. 835–842.

Armstrong, Natalie (2018). "Overdiagnosis and overtreatment as a quality problem: insights from healthcare improvement research". In: *BMJ quality & safety* 27.7. Publisher: BMJ Publishing Group Ltd, pp. 571–575.

Athey, Susan and Guido W. Imbens (2019). "Machine Learning Methods That Economists Should Know About". In: *Annual Review of Economics* 11.1. _eprint: https://doi.org/10.1146/annurev-economics-080217-053433, pp. 685–725.

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 (Nov. 2015). "Behavioral Hazard in Health Insurance". In: *The Quarterly Journal of Economics* 130.4, pp. 1623–1667.

Baim, D. S. and D. I. Simon (2006). "Complications and the optimal use of adjunctive pharmacology". In: *Grossman's cardiac catheterization, angiography, and intervention. Philadelphia: Lippincott Williams & Wilkins*, pp. 42–7.

Baker, Laurence C. (2001). "Managed care and technology adoption in health care: evidence from magnetic resonance imaging". In: *Journal of health economics* 20.3. Publisher: Elsevier, pp. 395–421.

Bavry, Anthony A., Dharam J. Kumbhani, Andrew N. Rassi, Deepak L. Bhatt, and Arman T. Askari (2006). "Benefit of early invasive therapy in acute coronary syndromes: a meta-analysis of contemporary randomized clinical trials". In: *Journal of the American College of Cardiology* 48.7, pp. 1319–1325.

Betsou, S., E. P. Efstathopoulos, D. Katritsis, K. Faulkner, and G. Panayiotakis (1998). "Patient radiation doses during cardiac catheterization procedures." In: *The British journal of radiology* 71.846, pp. 634–639.

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 (Apr. 2012). "Salience Theory of Choice Under Risk". In: *The Quarterly Journal of Economics*, qjs018.

— (2017). "Memory, attention, and choice". In: *National Bureau of Economic Research Working Paper*.

Brenner, David J., Richard Doll, Dudley T. Goodhead, Eric J. Hall, Charles E. Land, John B. Little, Jay H. Lubin, Dale L. Preston, R. Julian Preston, Jerome S. Puskin, Elaine Ron, Rainer K. Sachs, Jonathan M. Samet, Richard B. Setlow, and Marco Zaider (Nov. 2003). "Cancer risks attributable to low doses of ionizing radiation: Assessing what we really know". In: *Proceedings of the National Academy of Sciences* 100.24, pp. 13761–13766.

Brot-Goldberg, Zarek C., Amitabh Chandra, Benjamin R. Handel, and Jonathan T. Kolstad (2015). "What does a deductible do? the impact of cost-sharing on health care prices, quantities, and spending dynamics". In: *National Bureau of Economic Research Working Paper*.

Brown, Michael D., Stephen J. Wolf, Richard Byyny, Deborah B. Diercks, Seth R. Gemme, Charles J. Gerardo, Steven A. Godwin, Sigrid A. Hahn, Nicholas E. Harrison, Benjamin W. Hatten, Jason S. Haukoos, Amy Kaji, Heemun Kwok, Bruce M. Lo, Sharon E. Mace, Devorah J. Nazarian, Jean A. Proehl, Susan B. Promes, Kaushal H. Shah, Richard D. Shih, Scott M. Silvers, Michael D. Smith, Molly E. W. Thiessen, Christian A. Tomaszewski, Jonathan H. Valente, Stephen P. Wall, Stephen V. Cantrill, Jon Mark Hirshon, Travis Schulz, Rhonda R. Whitson, Christian A. Tomaszewski, David Nestler, Kaushal H. Shah, Amita Sudhir, and Michael D. Brown (Nov. 2018). "Clinical Policy: Critical Issues in the Evaluation and Management of Emergency Department Patients With Suspected Non–ST-Elevation Acute Coronary Syndromes". In: *Annals of Emergency Medicine* 72.5, e65–e106.

Camerer, Colin (2018). *Artificial Intelligence and Behavioral Economics*. NBER Chapters. National Bureau of Economic Research, Inc.

Chan, David C. and Jonathan Gruber (May 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 (Nov. 2019). *Selection with Variation in Diagnostic Skill: Evidence from Radiologists*. Working Paper 26467. Series: Working Paper Series. National Bureau of Economic Research.

Chandra, Amitabh, Jonathan Gruber, and Robin McKnight (Mar. 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.

Chandrasekar, Baskaran, Serge Doucet, Luc Bilodeau, Jacques Crepeau, Pierre deGuise, Jean Gregoire, Richard Gallo, Gilles Cote, Raoul Bonan, Michel Joyal, Gilbert Gosselin, Jean-François Tanguay, Ihor Dyrda, Marc Bois, and Andre Pasternac (2001). "Complications of cardiac catheterization in the current era: A single-center experience". In: *Catheterization and Cardiovascular Interventions* 52.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ccd.10 pp. 289–295.

Cobb, Leonard A. and W. Douglas Weaver (Jan. 1986). "Exercise: A risk for sudden death in patients with coronary heart disease". In: *Journal of the American College of Cardiology* 7.1, pp. 215–219.

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.

Einav, Liran, Amy Finkelstein, Tamar Oostrom, Abigail J. Ostriker, and Heidi L. Williams (2019). *Screening and selection: The case of mammograms*. Tech. rep. National Bureau of Economic Research.

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

Ely, Sora, Abhinav Chandra, Giselle Mani, Weiying Drake, Debbie Freeman, and Alexander T. Limkakeng (Feb. 2013). "Utility of Observation Units for Young Emergency Department Chest Pain Patients". In: *Journal of Emergency Medicine* 44.2. Publisher: Elsevier, pp. 306–312.

Fisher, Elliott S., David E. Wennberg, Thrse 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, p. 288.

Foy AJ, Liu G, Davidson WR, Jr, Sciamanna C, and Leslie DL (Mar. 2015). "Comparative effectiveness of diagnostic testing strategies in emergency department patients with chest pain: An analysis of downstream testing, interventions, and outcomes". In: *JAMA Internal Medicine* 175.3, pp. 428–436.

Friedman, Jerome H. (2001). "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5, pp. 1189–1232.

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

— (2017). *Behavioral inattention*. Tech. rep. National Bureau of Economic Research.

Ghassemi, Marzyeh, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits (2014). "Unfolding physiological state: Mortality modelling in intensive care units". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 75–84.

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.

Graber ML, Wachter RM, and Cassel CK (Sept. 2012). "Bringing diagnosis into the quality and safety equations". In: *JAMA* 308.12, pp. 1211–1212.

Greenberg, Jerome and Jonas B. Green (2014). "Over-testing: Why More Is Not Better". In: *The American Journal of Medicine* 5.127, pp. 362–363.

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.

Henry, Katharine E., David N. Hager, Peter J. Pronovost, and Suchi Saria (2015). "A targeted real-time early warning score (TREWScore) for septic shock". In: *Science translational medicine* 7.299, 299ra122–299ra122.

Hermann, Luke K., David H. Newman, W. Andrew Pleasant, Dhanadol Rojanasarntikul, Daniel Lakoff, Scott A. Goldberg, W. Lane Duvall, and Milena J. Henzlova (2013). "Yield

of routine provocative cardiac testing among patients in an emergency department–based chest pain unit". In: *JAMA internal medicine* 173.12, pp. 1128–1133.

Hill, J. D., J. R. Hampton, and J. R. Mitchell (Apr. 1978). "A randomised trial of home-versus-hospital management for patients with suspected myocardial infarction". In: *Lancet (London, England)* 1.8069, pp. 837–841.

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

Jeremias, Allen and C. Michael Gibson (May 2005). "Narrative Review: Alternative Causes for Elevated Cardiac Troponin Levels when Acute Coronary Syndromes Are Excluded". In: *Annals of Internal Medicine* 142.9. Publisher: American College of Physicians, pp. 786–791.

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

Katzenschlager, Reinhold, Ara Ugurluoglu, Ali Ahmadi, M. Hülsmann, Renate Koppensteiner, Elisabeth Larch, Thomas Maca, Erich Minar, A. Stümpflen, and Herbert Ehringer (1995). "Incidence of pseudoaneurysm after diagnostic and therapeutic angiography." In: *Radiology* 195.2, pp. 463–466.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2017). "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.

Kohn, Linda T., Janet Corrigan, and Molla S. Donaldson (2000). *To err is human: building a safer health system*. Vol. 6. National academy press Washington, DC.

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.

Lee, Thomas H., Gregory W. Rouan, Monica C. Weisberg, Donald A. Brand, Denise Acampora, Carol Stasiulewicz, Jay Walshon, George Terranova, Louis Gottlieb, Beth Goldstein-Wayne, David Copen, Karen Daley, Allan A. Brandt, John Mellors, Rita Jakubowski, E. Francis Cook, and Lee Goldman (Aug. 1987). "Clinical characteristics and natural history of patients with acute myocardial infarction sent home from the emergency room". In: *American Journal of Cardiology* 60.4, pp. 219–224.

Liao, Joshua M., Lee A. Fleisher, and Amol S. Navathe (Sept. 2016). "Increasing the Value of Social Comparisons of Physician Performance Using Norms". In: *JAMA* 316.11, pp. 1151–1152.

Litt, Harold I., Constantine Gatsonis, Brad Snyder, Harjit Singh, Chadwick D. Miller, Daniel W. Entrikin, James M. Leaming, Laurence J. Gavin, Charissa B. Pacella, and Judd E. Hollander (Apr. 2012). "CT Angiography for Safe Discharge of Patients with Possible Acute Coronary Syndromes". In: *New England Journal of Medicine* 366.15, pp. 1393–1403.

Loewenstein, George, Kevin G. Volpp, and David A. Asch (Apr. 2012). "Incentives in Health: Different Prescriptions for Physicians and Patients". In: *JAMA* 307.13, pp. 1375–1376.

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, Laura A. Demopoulos, Peter M. DiBattiste, Christopher P. Cannon, William S. Weintraub, and for the TACTICS-TIMI 18 Investigators (Oct. 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.

Mather, H G, D C Morgan, N G Pearson, K L Read, D B Shaw, G R Steed, M G Thorne, C J Lawrence, and I S Riley (Apr. 1976). "Myocardial infarction: a comparison between home and hospital care for patients." In: *British Medical Journal* 1.6015, pp. 925–929.

Mettler, Fred A., Walter Huda, Terry T. Yoshizumi, and Mahadevappa Mahesh (July 2008). "Effective doses in radiology and diagnostic nuclear medicine: a catalog". In: *Radiology* 248.1, pp. 254–263.

Miotto, Riccardo, Li Li, Brian A. Kidd, and Joel T. Dudley (2016). "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records". In: *Scientific reports* 6, p. 26094.

Morden, Nancy E., Carrie H. Colla, Thomas D. Sequist, and Meredith B. Rosenthal (Feb. 2014). "Choosing Wisely — The Politics and Economics of Labeling Low-Value Services". In: *The New England journal of medicine* 370.7, pp. 589–592.

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 (2016). "Machine Learning: An Applied Econometric Approach". In: *Journal of Economic Perspectives* Forthcoming.

— (2017). "Machine learning: an applied econometric approach". In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Neumann, Peter J., Joshua T. Cohen, and Milton C. Weinstein (Aug. 2014). "Updating Cost-Effectiveness — The Curious Resilience of the $50,000-per-QALY Threshold". In: *New England Journal of Medicine* 371.9, pp. 796–797.

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

Newman-Toker, David E., Ernest Moy, Ernest Valente, Rosanna Coffey, and Anika L. Hines (2014). "Missed diagnosis of stroke in the emergency department: a cross-sectional analysis of a large population-based sample". In: *Diagnosis* 1.2, pp. 155–166.

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: *BMJ* 356, j239.

O'Sullivan, Jack W., Ali Albasri, Brian D. Nicholson, Rafael Perera, Jeffrey K. Aronson, Nia Roberts, and Carl Heneghan (Feb. 2018). "Overtesting and undertesting in primary care: a systematic review and meta-analysis". In: *BMJ Open* 8.2. Publisher: British Medical Journal Publishing Group Section: Epidemiology, e018557.

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. American Medical Informatics Association, p. 1109.

Poldervaart, J. M., M. Langedijk, B. E. Backus, I. M. C. Dekker, A. J. Six, P. A. Doevendans, A. W. Hoes, and J. B. Reitsma (Jan. 2017). "Comparison of the GRACE, HEART and TIMI score to predict major adverse cardiac events in chest pain patients at the emergency department". In: *International Journal of Cardiology* 227, pp. 656–661.

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". In: *Arch Intern Med* 172.19, pp. 1506–1509.

Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane (2019). "Machine learning in medicine". In: *New England Journal of Medicine* 380.14, pp. 1347–1358.

Rajkomar, Alvin, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, and Mimi Sun (2018). "Scalable and accurate deep learning with electronic health records". In: *NPJ Digital Medicine* 1.1, p. 18.

Redberg, Rita F. (2015). "Stress Testing in the Emergency Department: Not Which Test but Whether Any Test Should Be Done". In: *JAMA internal medicine* 175.3, pp. 436–436.

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. 358–360.

Rich, Michael W. and Charles A. Crecelius (June 1990). "Incidence, Risk Factors, and Clinical Course of Acute Renal Insufficiency After Cardiac Catheterization in Patients 70 Years of Age or Older: A Prospective Study". In: *Archives of Internal Medicine* 150.6, pp. 1237–1242.

Ridker, Paul M, Eleanor Danielson, Francisco A.H. Fonseca, Jacques Genest, Antonio M. Gotto, John J.P. Kastelein, Wolfgang Koenig, Peter Libby, Alberto J. Lorenzatti, Jean G. MacFadyen, Børge G. Nordestgaard, James Shepherd, James T. Willerson, and Robert J. Glynn (Nov. 2008). "Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated C-Reactive Protein". In: *New England Journal of Medicine* 359.21, pp. 2195–2207.

Rokos, Ivan C., William J. French, Amal Mattu, Graham Nichol, Michael E. Farkouh, James Reiffel, and Gregg W. Stone (2010). "Appropriate cardiac cath lab activation: optimizing electrocardiogram interpretation and clinical decision-making for acute ST-elevation myocardial infarction". In: *American heart journal* 160.6. Publisher: Elsevier, pp. 995–1003.

Rozanski, Alan, Heidi Gransar, Sean W. Hayes, James Min, John D. Friedman, Louise E.J. Thomson, and Daniel S. Berman (Mar. 2013). "Temporal Trends in the Frequency of

Inducible Myocardial Ischemia During Cardiac Stress Testing1991 to 2009". In: *Journal of the American College of Cardiology* 61.10, pp. 1054–1065.

Schor S, Behar S, Modan B, Barell V, Drory J, and Kariv I (Aug. 1976). "Disposition of presumed coronary patients from an emergency room: A follow-up study". In: *JAMA* 236.8, pp. 941–943.

Schulman, K. A., J. A. Berlin, W. Harless, J. F. Kerner, S. Sistrunk, B. J. Gersh, R. Dubé, C. K. Taleghani, J. E. Burke, S. Williams, and others (1999). "The effect of race and sex on physicians' recommendations for cardiac catheterization." In: *The New England journal of medicine* 340.8, p. 618.

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.

Sharp, Adam L., Benjamin Broder, and Benjamin C Sun (Apr. 2018). *HEART Score Improves ED Care for Low-Risk Chest Pain.*

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.

Singh, Hardeep (2013). "Diagnostic errors: Moving beyond 'no respect'and getting ready for prime time". In: *BMJ quality & safety* 22.10, pp. 789–792.

Stripe, Benjamin, Stephen Rechenmacher, Daniel Jurewitz, Cody Lee, and Saul Schaefer (Dec. 2013). "The Diagnostic Yield of Cardiac Catheterization in Low-Risk Troponinemia". In: *JAMA Internal Medicine* 173.22, p. 2088.

Swap CJ and Nagurney JT (Nov. 2005). "Value and limitations of chest pain history in the evaluation of patients with suspected acute coronary syndromes". In: *JAMA* 294.20, pp. 2623–2629.

Tang, Eng Wei, Cheuk-Kit Wong, and Peter Herbison (2007). "Global Registry of Acute Coronary Events (GRACE) hospital discharge risk score accurately predicts long-term mortality post acute coronary syndrome". In: *American heart journal* 153.1, pp. 29–35.

Than, Martin, Louise Cullen, Christopher M Reid, Swee Han Lim, Sally Aldous, Michael W Ardagh, W Frank Peacock, William A Parsonage, Hiu Fai Ho, Hiu Fai Ko, Ravi R Kasliwal, Manish Bansal, Sunarya Soerianata, Dayi Hu, Rongjing Ding, Qi Hua, Kang Seok-Min, Piyamitr Sritara, Ratchanee Sae-Lee, Te-Fa Chiu, Kuang-Chau Tsai, Fang-Yeh Chu, Wei-Kung Chen, Wen-Han Chang, Dylan F Flaws, Peter M George, and A Mark Richards (Mar. 2011). "A 2-h diagnostic protocol to assess patients with chest pain symptoms in the Asia-Pacific region (ASPECT): a prospective observational validation study". In: *The Lancet* 377.9771, pp. 1077–1084.

Than, Martin, Mel Herbert, Dylan Flaws, Louise Cullen, Erik Hess, Judd E. Hollander, Deborah Diercks, Michael W. Ardagh, Jeffery A. Kline, Zea Munro, and Allan Jaffe (July 2013). "What is an acceptable risk of major adverse cardiac event in chest pain patients soon after discharge from the Emergency Department?: A clinical survey". In: *International Journal of Cardiology* 166.3, pp. 752–754.

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

— (1981). "The framing of decisions and the psychology of choice". In: *Science* 211.4481, pp. 453–458.

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.

Welch, H. Gilbert, Lisa Schwartz, and Steve Woloshin (2011). *Overdiagnosed: making people sick in the pursuit of health*. Beacon Press.

Wyman, R. Michael, Robert D. Safian, Valerie Portway, John J. Skillman, Raymond G. Mckay, and Donald S. Baim (Dec. 1988). "Current complications of diagnostic and therapeutic cardiac catheterization". In: *Journal of the American College of Cardiology* 12.6, pp. 1400–1406.

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
N Patients	130,059	6,088	123,971
N Visits	246,874	7,320	239,554
Demographics			
Age, mean	42	58	42
	(0.033)	(0.146)	(0.033)
Female	0.611	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	0.432
	(<0.001)	(0.006)	(0.001)
Risk factors			
Past Heart Disease	0.121	0.391	0.113
	(<0.001)	(0.006)	(<0.001)
Diabetes	0.142	0.294	0.137
	(<0.001)	(0.005)	(<0.001)
Hypertension	0.251	0.513	0.243
	(<0.001)	(0.006)	(<0.001)
Cholesterol	0.162	0.417	0.155
	(<0.001)	(0.006)	(<0.001)
Any Risk Factor	0.36	0.625	0.351
	(<0.001)	(0.006)	(<0.001)
Triage Shifts			
Number of Shifts	5,925		
Patients per Shift	42		

Notes: Hospital sample descriptive statistics (mean (SE)). Includes all consecutive emergency visits from 2010-2015, excluding those over 80 years old, those with prior life-limiting illness (e.g., cancer, dementia, palliative or hospice care), those with cardiac procedure in previous 30 days, and those who died in the ER. We also exclude untested patients with evidence of heart attack in ER.

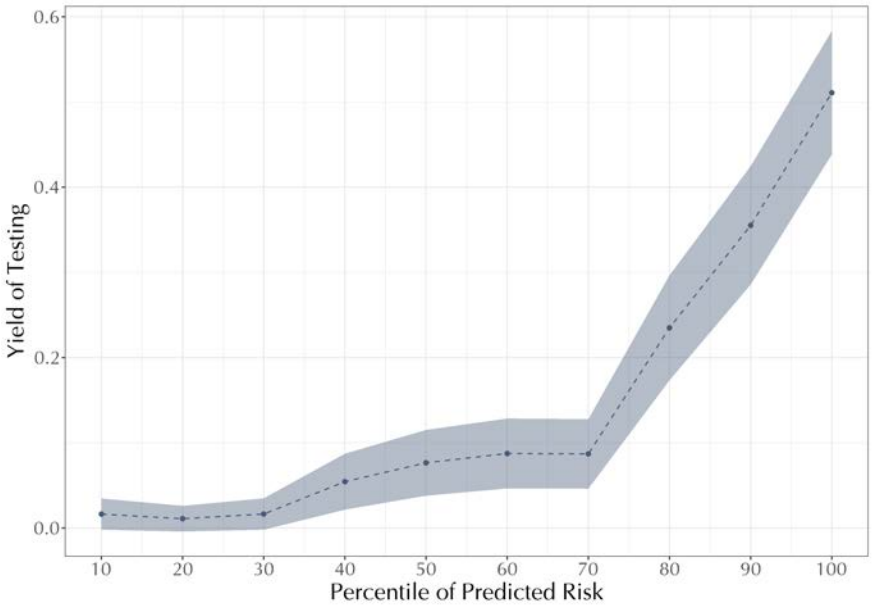
Table 2: Summary Statistics: Physician Choices and Patient Outcomes

	All	Tested	Untested
Physician Suspicion for Heart Attack			
ECG in ER	0.298	0.975	0.277
	(<0.001)	(0.002)	(<0.001)
Troponin Test in ER	0.112	0.792	0.092
	(<0.001)	0.005	(<0.001)
Heart Attack Diagnosis	0.032	0.403	0.019
	(<0.001)	(0.005)	(<0.001)
Positive Troponin Test in ER	0.052	0.288	0.044
	(<0.001)	(0.005)	(<0.001)
Mean Troponin, if Positive	-	0.72	-
	-	(0.005)	-
Testing Rate (10 Day)			
Overall Test Rate	0.030	-	-
	(<0.001)	-	-
Catheterization Rate	0.012	-	-
	(<0.001)	-	-
Stress Test Rate	0.020	-	-
	(<0.001)	-	-
Outcomes of Testing (10 Day)			
Overall Yield	-	0.146	-
	-	(0.004)	-
Yield: Stenting	-	0.129	-
	-	(0.004)	-
Yield: Open-heart Surgery	-	0.018	-
	-	(0.002)	-
Adverse Events			
Any Adverse Event (30 Day)	0.019	0.261	0.012
	(<0.001)	(0.005)	(<0.001)
Diagnosed Heart Attack	0.016	0.253	0.009
or Arrythmia (30 Day)	(<0.001)	(0.005)	(<0.001)
Death (30 Day)	0.004	0.017	0.004
	(<0.001)	(0.002)	(<0.001)
Any Adverse Event (31-365 Day)	0.031	0.064	0.03
	(<0.001)	(0.003)	(<0.001)

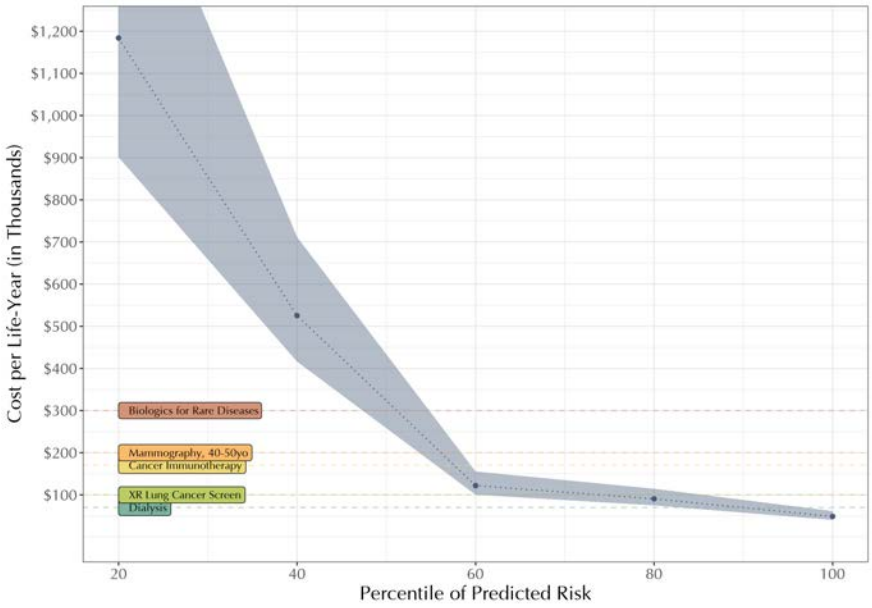
Notes: Key statistics in our hospital sample (mean (SE)), related to physicians' suspicion for heart attack during the ER visit, testing rate and type, outcome of testing, and subsequent major adverse cardiac events.

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) 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 from the literature.

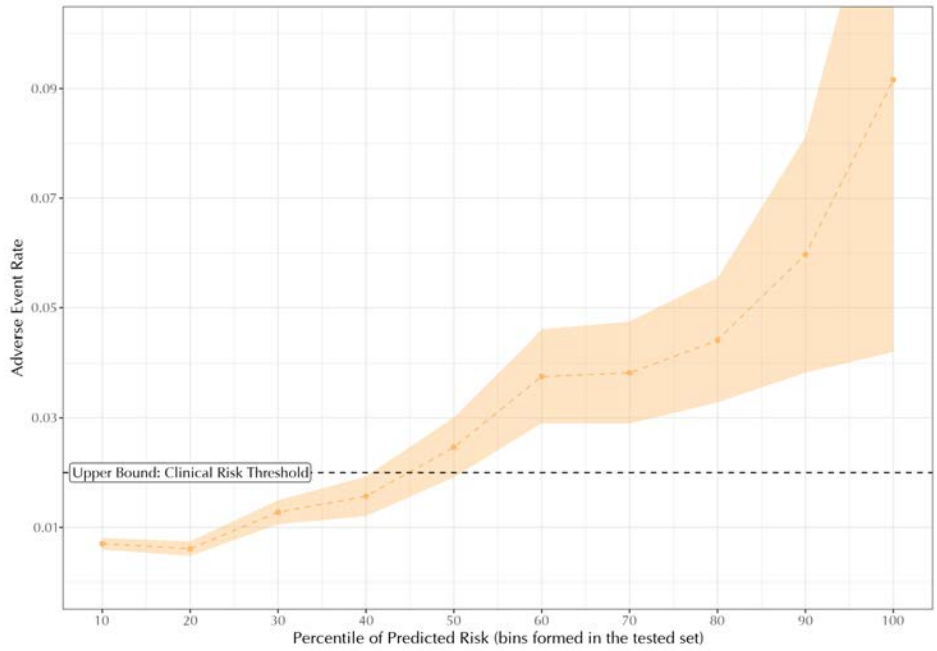
Table 3: Realized Yield, Cost-Effectiveness, and Testing Rate

	Yield Rate (SE) (1)	Cost-Effectiveness ($) (Lower–Upper Bound) (2)	Test Rate (SE) (3)
Full Sample	0.145	86,683	0.029
	(0.008)	(71,652-109,696)	(<0.001)
By Risk Bin			
1	0.014	1,183,936	0.01
	(0.006)	(901,105-1,725,532)	(0.001)
2	0.035	525,279	0.024
	(0.01)	(416,207-711,819)	(0.001)
3	0.082	121,935	0.068
	(0.014)	(100,286-155,502)	(0.003)
4	0.161	90,629	0.112
	(0.019)	(74,960-114,581)	(0.005)
5	0.433	48,831	0.38
	(0.026)	(40,502-61,473)	(0.016)
N	1,834	1,834	61,821

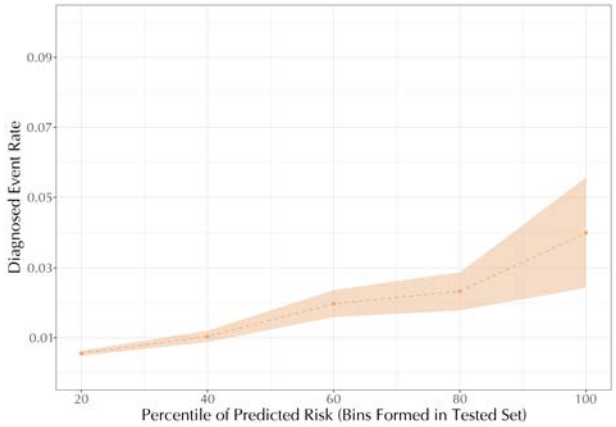
Notes: Realized yield of testing (1) and cost-effectiveness of testing (2) in the tested, and test rate across all visits (3), by bin of predicted risk. Bins are quintiles of risk, defined in the tested population (so bins are equally sized in Columns (1) and (2), but not in (3)).

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

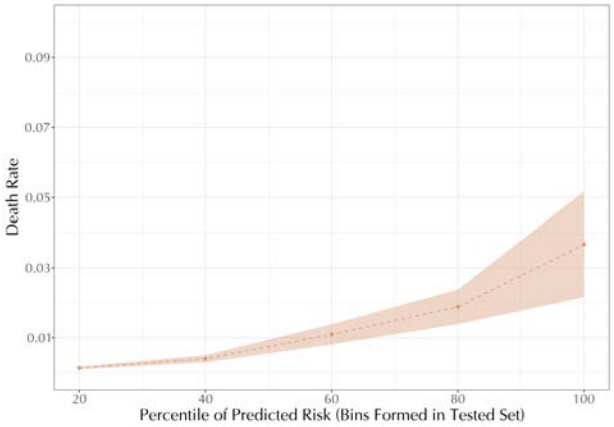
(a) Any Adverse Event



(b) Diagnosed Heart Attack or Arrythmia



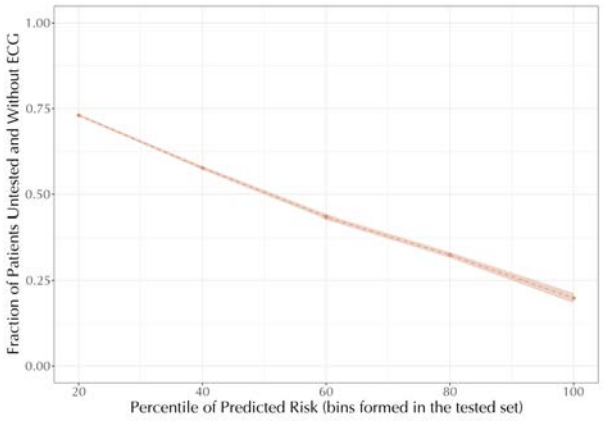
(c) Death



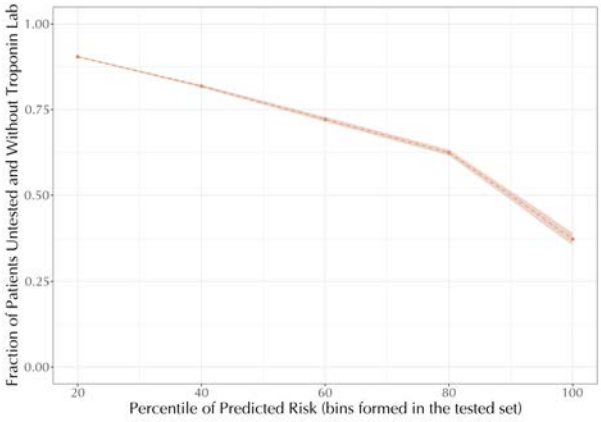
Notes: Rate of adverse events and death over the 30 days following visits (y-axis) among untested patients, by bin of predicted risk (x-axis). Risk bins are formed as deciles of predicted risk in the tested, for comparison (so bins are not equally sized). Top panel (a): combined rate of both diagnosed events (heart attack confirmed with laboratory biomarkers, and cardiac arrest) and death (measured by linking to Social Security data). The horizontal line shows the 2% threshold above which testing is recommended; 14% of the untested (drawn from the top 6 bins) have a rate significantly above 2%. Top of the highest 95% CI truncated. Bottom panels separately show components of total adverse event rate: (b) diagnosed adverse events and (c) death, with bins formed as quintiles of predicted risk in the tested (because outcomes are less frequent).

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

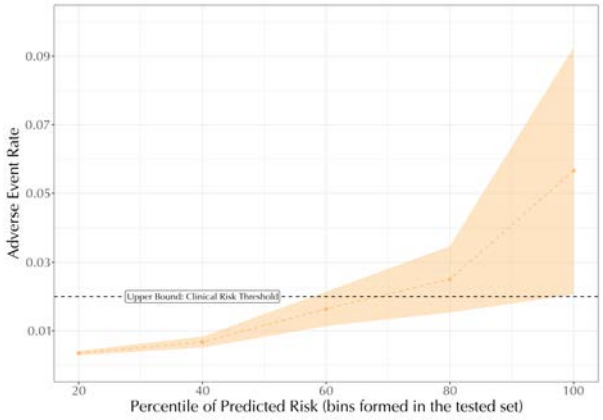
(a) Fraction Untested Without an ECG



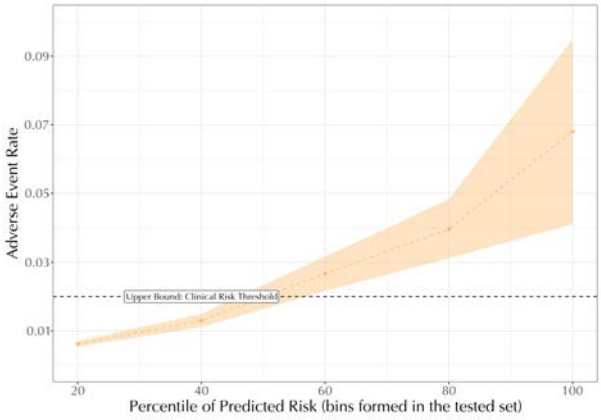
(b) Fraction Untested Without a Troponin Test



(c) Adverse Events, Patients Without ECG



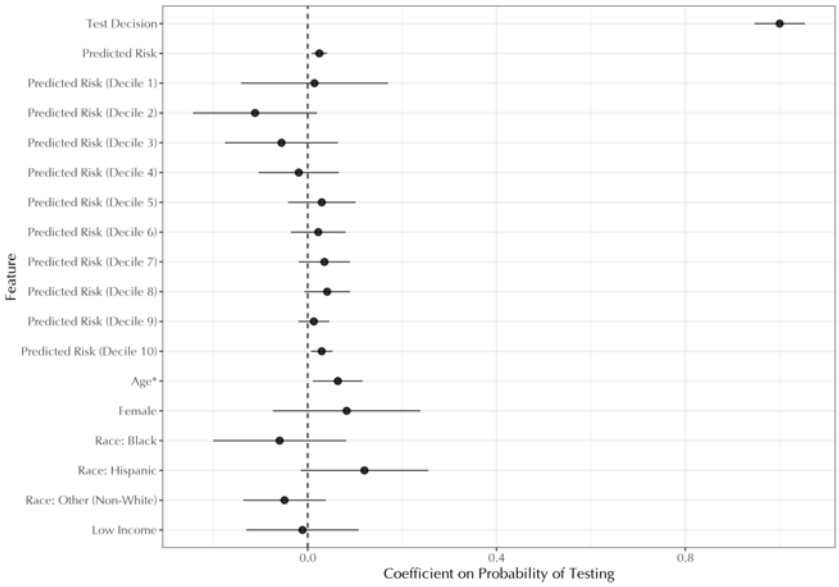
(d) Adverse Events, Patients Without Troponin



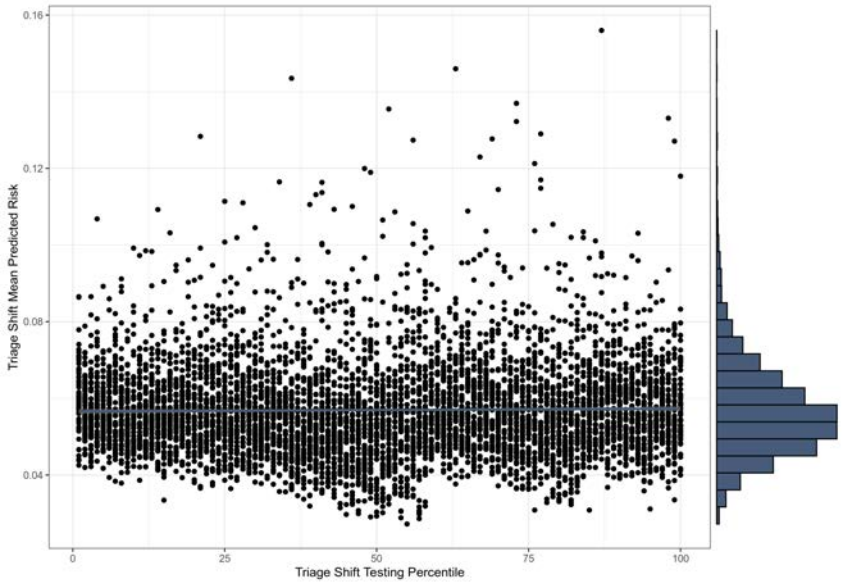
Notes: Top panels: fraction of untested patients in whom physicians do not appear to suspect heart attack, based on (a) the lack of an electrocardiogram (ECG), or (b) lack of troponin testing (a laboratory study that can indicate heart attack). Rates are shown by risk bins, which are formed as deciles of predicted risk in the tested for comparison (so bins are not equally sized here). Bottom panels: rate of major adverse cardiac events and death over the 30 days following emergency visits (y-axis) among these untested and unsuspected patients, by bin of predicted risk (x-axis). Panel (c) shows rates among untested patients without an ECG, panel (d) shows rates among untested patients without a troponin (right). 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: Balance checks for a 'natural experiment,' in which patients are tested at higher or lower rates (conditional on time of arrival and predicted risk), based on the triage team working when they arrive. Panel (a): Results of regressions of testing and pre-triage variables on triage testing rate (predicted by shift testing random effects), to check for balance on observable factors. Panel (b): average predicted risk of patients presenting on a given triage shift (y-axis) vs. the triage shift testing rate (x-axis, percentile of shift testing random effect). Each point represents one of 5,925 shifts in our entire dataset, and the density plot on the shows distribution of mean risk.

Table 4: Balance on Realized Yield of Testing by Shift Testing Rate

	Yield			
	(1)	(2)	(3)	(4)
Testing				
Shift Effect	0.006		0.005	
	(0.007)		(0.009)	
Shift Q2		−0.060		0.020
		(0.042)		(0.057)
Shift Q3		−0.036		−0.023
		(0.038)		(0.053)
Shift Q4		−0.026		−0.006
		(0.037)		(0.051)
Testing by Predicted Risk				
Risk × Shift Effect			0.007	
			(0.042)	
Risk × Shift Q2				−0.403**
				(0.170)
Risk × Shift Q3				−0.015
				(0.156)
Risk × Shift Q4				−0.051
				(0.152)
Risk Controls	Yes	Yes	Yes	Yes
Observations	1,830	1,830	1,830	1,830

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

Notes: Results of regressions to check for unobservable differences in patient risk across high- vs. low-testing triage shifts. Each column shows a regression of yield of testing (in the tested) on shift testing rate (measured by random effects for testing). We assume that if higher testing rates were driven by unobservable risk differences, patients tested in higher-testing shifts would have higher yield when tested.

Table 5: Average Effect of Testing on Long-Term Adverse Events

Testing Effect (Linear)	(1) Adverse Event (31-365 days)	(2) Diagnosed Event (31-365 days)	(3) Death (31-365 days)	(4) Death (365 days)
Shift Effect	−0.038	−0.007	−0.049**	−0.022
	(0.036)	(0.028)	(0.025)	(0.028)
Risk Control	Yes	Yes	Yes	Yes
Observations	213,484	213,484	213,484	213,484
R^2	0.010	0.003	0.012	0.021
Testing Effect (Quartiles)	Adverse Event (31-365 days)	Diagnosed Event (31-365 days)	Death (31-365 days)	Death (365 days)
Shift Q2	−0.040	0.015	−0.084	−0.086
	(0.100)	(0.079)	(0.069)	(0.078)
Shift Q3	0.140	0.160**	−0.021	−0.0001
	(0.100)	(0.079)	(0.069)	(0.078)
Shift Q4	−0.010	0.055	−0.116*	−0.068
	(0.100)	(0.080)	(0.069)	(0.078)
Risk Controls	Yes	Yes	Yes	Yes
Observations	213,484	213,484	213,484	213,484
R^2	0.006	0.001	0.008	0.015
Outcome Rates (%)	2.761	1.712	1.297	1.678

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

Notes: Average effect of testing on adverse outcomes in the year after ED visits, estimated in a 'natural experiment' in which patients are as-if-randomly assigned to higher or lower testing rates. The top panel uses a linear measure of shift testing rate, and the bottom panel uses a set of indicators for quartile of shift rate. Both specifications control for patient risk (not shown).

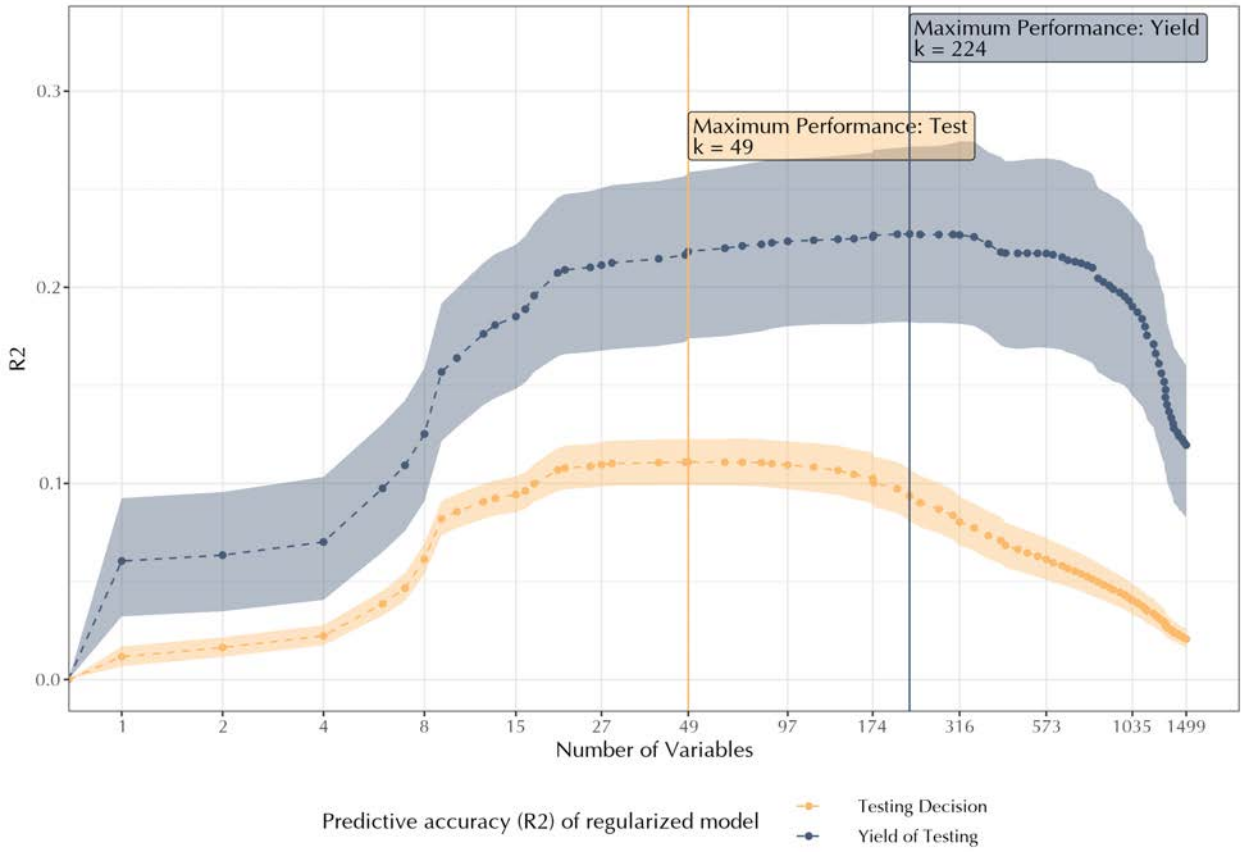
Table 6: Effect of Testing on Long-Term Adverse Events By Predicted Risk

Risk Quintiles by Testing (Linear)	(1) Adverse Event (31-365 days)	(2) Diagnosed Event (31-365 days)	(3) Death (31-365 days)	(4) Death (0-365 days)
Testing	−0.037	−0.037	−0.028	−0.024
	(0.061)	(0.049)	(0.042)	(0.048)
Risk Q2 × Testing	0.070	0.083	0.010	0.032
	(0.084)	(0.066)	(0.058)	(0.065)
Risk Q3 × Testing	0.085	0.128	−0.019	0.011
	(0.102)	(0.081)	(0.070)	(0.080)
Risk Q4 × Testing	−0.316**	−0.129	−0.201*	−0.084
	(0.153)	(0.121)	(0.105)	(0.119)
Risk Q5 × Testing	−1.373***	−1.093***	−0.432**	−0.460**
	(0.275)	(0.219)	(0.190)	(0.215)
Risk Controls	Yes	Yes	Yes	Yes
R^2	0.010	0.003	0.012	0.021
Risk (Linear) by Testing (Quartiles)	Adverse Event (31-365 days)	Diagnosed Event (31-365 days)	Death (31-365 days)	Death (0-365 days)
Shift Q2	0.242*	0.208**	0.040	0.026
	(0.130)	(0.103)	(0.090)	(0.102)
Shift Q3	0.524***	0.375***	0.159*	0.160
	(0.129)	(0.102)	(0.089)	(0.100)
Shift Q4	0.407***	0.289***	0.094	0.141
	(0.129)	(0.102)	(0.089)	(0.100)
Risk × Shift Q2	−6.027***	−4.151***	−2.636**	−2.379*
	(1.807)	(1.434)	(1.246)	(1.409)
Risk × Shift Q3	−8.223***	−4.599***	−3.850***	−3.431**
	(1.739)	(1.380)	(1.200)	(1.357)
Risk × Shift Q4	−8.908***	−4.974***	−4.473***	−4.462***
	(1.720)	(1.364)	(1.186)	(1.341)
Risk Controls	Yes	Yes	Yes	Yes
R^2	0.006	0.001	0.008	0.015
Observations	213,484	213,484	213,484	213,484
Outcome Rates (%)	2.761	1.712	1.297	1.678

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

Notes: Effect of testing on adverse outcomes, as in Table 5, but where effects vary by predicted risk. The top panel uses linear testing rate, interacted with risk quintiles; the bottom panel uses testing quartile, interacted with linear risk. Both control for patient risk.

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^*).

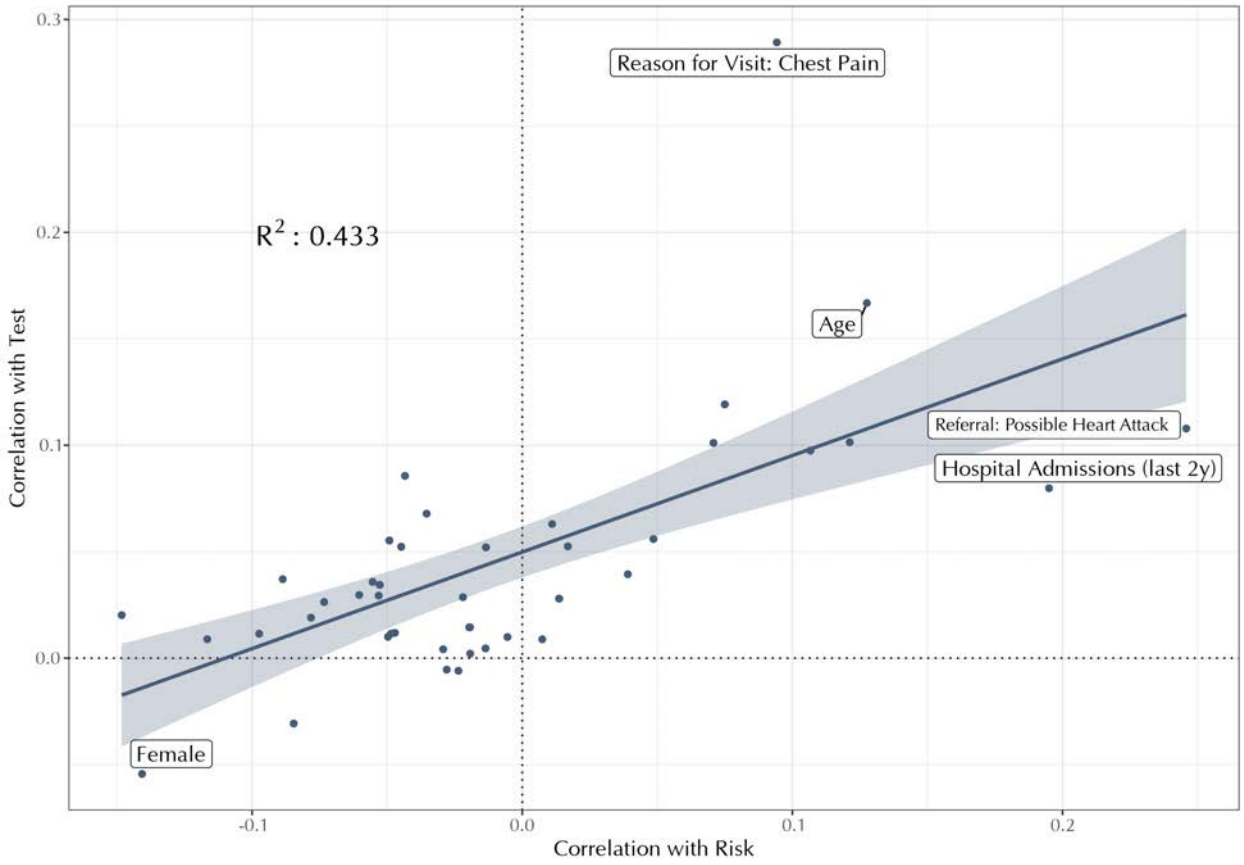
Table 7: Test for Physician Boundedness

	Test		Yield	
	(1)	(2)	(3)	(4)
Predicted Risk, Simple ($k = 49$)	1.357***	1.358***	1.528***	1.319***
	(0.015)	(0.016)	(0.068)	(0.081)
Incremental Risk, Complex ($k = 224$)		−0.005		1.099***
		(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: Correlation of Variables in Simple Risk Model with Physician's Testing Decision and Patient Risk



Notes: For the simple risk model of complexity $k_h^* = 49$ that best predicts doctors' 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. Some outlier points are labeled.

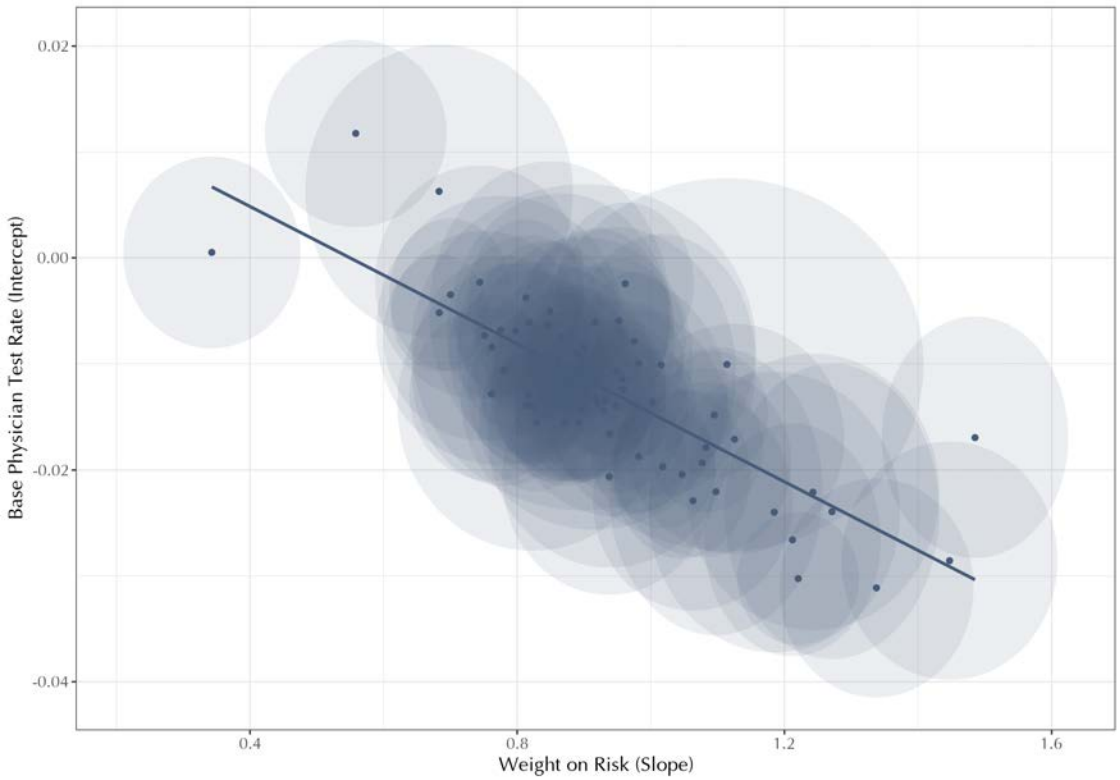
Table 8: Symptom Salience and Representativeness

	Test			
	(1)	(2)	(3)	(4)
Predicted Risk, Full	0.888***	0.462***	0.459***	0.105*
	(0.051)	(0.053)	(0.053)	(0.055)
Predicted Risk, Subsets of Inputs				
Symptoms		0.779***	0.801***	0.003
		(0.024)	(0.024)	(0.041)
Representative Symptoms				1.431***
				(0.061)
Demographics			0.150***	
			(0.018)	
Prior Diagnoses			0.083***	
			(0.012)	
Prior Procedures			−0.013	
			(0.014)	
Prior Labs			−0.060***	
			(0.012)	
Prior Vital Signs			−0.078***	
			(0.012)	
Medications			−0.070***	
			(0.015)	
Constant	0.122***	0.177***	0.176 ***	0.229***
	(0.028)	(0.028)	(0.028)	(0.028)
Risk Controls	Yes	Yes	Yes	Yes
Observations	61,821	61,821	61,821	61,821
R^2	0.111	0.126	0.134	0.131

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

Notes: Column (1) shows a simple regression of testing on our usual predicted risk. Column (2) adds a predictor of risk, formed in the same way as our usual predicted risk, but using only symptoms as inputs. Column (3) adds additional risk predictors to the regression in Column (2), formed using the various other types of inputs. Column (4) adds another version of symptom-based predicted risk, formed solely from the subset of 9 *representative* symptoms, to the regression in Column (2). All models control for non-linear risk terms (not shown). Appendix Table 18 shows results of similar regressions to predict yield of testing instead of the testing decision; it verifies that that none of the predicted risk variables based on subsets of inputs are additionally useful for predicting yield.

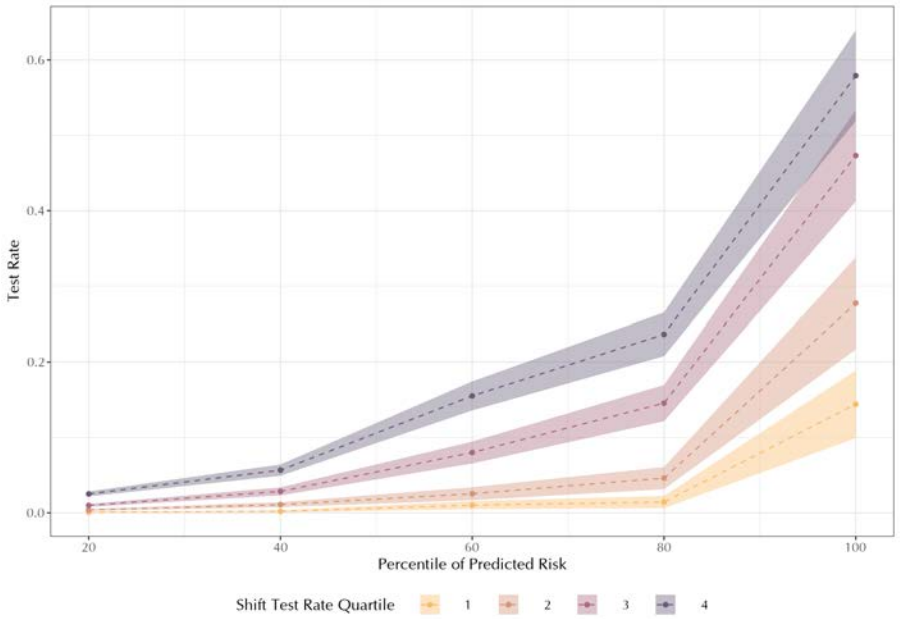
Figure 7: Physician Testing Propensity and Skill



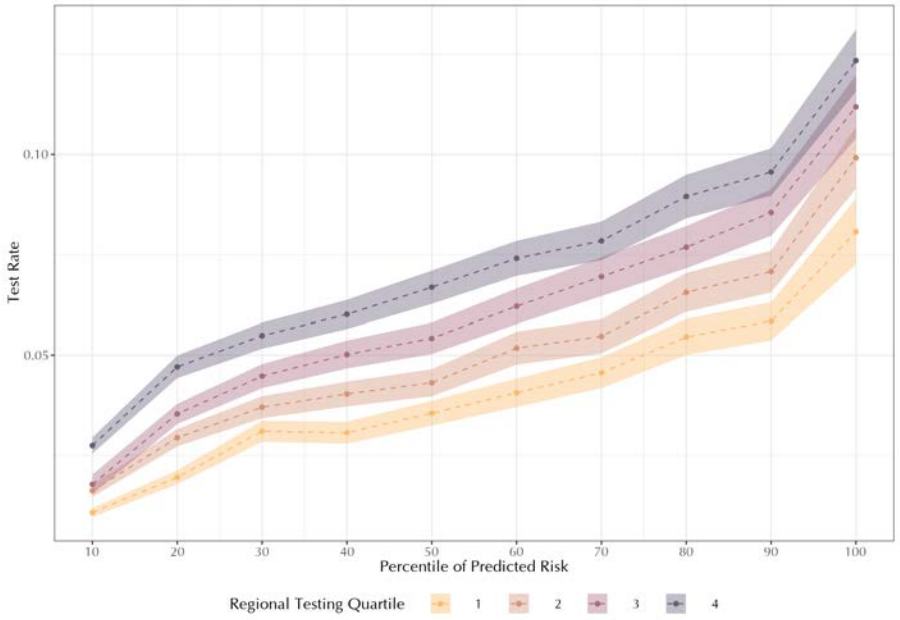
Notes: Results from a regression of the testing decision on predicted risk, with physician-specific slopes and intercepts for the 70 physicians with at least 500 visits in our sample. The slope on predicted risk, a measure of skill, is on the x-axis, and the intercept, a measure of testing propensity, is shown on the y-axis. Each point is one physician, and the ovals show the 95% confidence interval of each parameter. The line of best fit is also shown.

Figure 8: 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 (conditional on time of arrival and predicted risk), 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.