NBER WORKING PAPER SERIES

A MACHINE LEARNING APPROACH TO LOW-VALUE HEALTH CARE: WASTED TESTS, MISSED HEART ATTACKS AND MIS-PREDICTIONS

Sendhil Mullainathan
Ziad Obermeyer

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

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2019, Revised August 2020

Previously circulated as "Who is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error." Authors contributed equally. We acknowledge financial support from the Pershing Square Fund for Research on the Foundations of Human Behavior, grant DP5 OD012161 from the Office of the Director of the National Institutes of Health, and grant P01 AG005842 from the National Institute on Aging. We are deeply grateful to Advik Shreekumar, as well as Adam Baybutt, Brent Cohn, Christian Covington, Shreyas Lakhtakia, Katie Lin, Ruchi Mahadeshwar, Jasmeet Samra, Cassidy Shubatt, and Aly Valliani, for outstanding research assistance; and to Amitabh Chandra, Xavier Gabaix, Jon Kolstad, Suchi Saria, Andrei Shleifer and Richard Thaler for very helpful feedback on a draft. We are also appreciative of seminar participants at several institutions for their thoughtful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

A Machine Learning Approach to Low-Value Health Care: Wasted Tests, Missed Heart Attacks and Mis-Predictions
Sendhil Mullainathan and Ziad Obermeyer
NBER Working Paper No. 26168
August 2019, Revised August 2020
JEL No. C55,D8,D84,D9,I1,I13

ABSTRACT

We use machine learning to better characterize low-value health care and the decisions that produce it. We focus on costly tests, specifically for heart attack (acute coronary syndromes). A test is only useful if it yields new information, so efficient testing is grounded in accurate prediction of test outcomes. Physician testing decisions can therefore be benchmarked against tailored algorithmic predictions, which provide a more precise way to study low-value care than the usual approach—looking at average test yield. Implemented in a large national sample, this procedure reveals significant over-testing: 52.6%of high-cost tests for heart attack are wasted. At the same time, it also reveals significant under-testing: many patients with predictably high risk go untested, then experience frequent adverse cardiac events including death in the next 30 days. At standard clinical thresholds, these event rates suggest that testing these patients would indeed have been highly cost-effective. Of the potential welfare gains from more efficient testing, 42.8% would come from addressing under-use. Existing policy levers, however, appear too blunt a tool to address both over- and under-use inefficiencies. We find that they cut testing across the board, for low-risk (reducing over-use) and high-risk patients (exaggerating under-use). Finally, we uncover two behavioral mechanisms for physician testing errors: (i) bounded rationality, in which physicians use an overly narrow set of variables, but make effective use of that set; and (ii) representativeness, in which they over-weight how "representative" heart attack is for a patient, above and beyond the conditional probability. Together, these results suggest the need for models of low-value care that incorporate mis-prediction so as to account for both over- and under-testing.

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 inefficiency of the US health care system is plain to see: relative to comparable countries, spending is much higher and life expectancy is far lower (Papanicolas, Woskie, and Jha 2018). Pinpointing the sources of inefficiency, however, is more challenging. For example, over-testing—e.g., wasteful MRIs for back pain, CT scans for headache, or routine screening for cancer—is a commonly cited culprit (Morden et al. 2014). To empirically measure establish over-testing, we typically must rely on the rate at which it identifies a problem. Average test yield, though, is a coarse metric, unable to quantify the extent whether only a few or nearly all tests are wasted. More fundamentally, averages can reveal little about underlying mechanisms since physicians make testing decisions one patient at a time. Here we suggest that machine learning, by providing tailored, patient-specific risk estimates, allows for a more granular way to quantify the efficiency of medical decisions.

Machine learning is particularly applicable because the decision to test implicitly involves forming a prediction (Kleinberg et al. 2015). Imagine a patient arriving to the emergency department complaining of chest pain. A physician sees the patient and worries about heart attack (acute coronary syndromes). The immediately available tests are inconclusive, so she contemplates a costly and potentially invasive test. There are good reasons to do the test: if it is positive, she has caught heart attack early and can deploy highly cost-effective treatments that reduce mortality and long-term complications. But there are also reasons not to: if the test is negative, it will have been a waste, both in dollars and in potential risk to patient. In weighing the tradeoff, the key unknown is the probability that the test will be positive. Efficient testing can thus be reframed as accurate prediction. And by forming a predictor of test outcome, we can assess the efficiency of testing decisions.

We build on this insight to empirically study the specific case of heart attack, and in particular the decision to stress test or catheterize emergency patients. In a sample of nationwide Medicare claims data from 2009-13, we build an ensemble model (consisting of gradient boosted trees and penalized linear models) to predict whether patients tested for heart attack[1] will end up being treated for heart attack. The model, formed in a randomly selected training set, includes as inputs only information available at the time of the testing decision. In an independent hold-out set, we compare these algorithmic predictions to physician decisions.

We find a great deal of over-testing. Using standard cost effectiveness methodology and assumptions—i.e., weighing the cost of the test against the life-years it saves[2]—we can calculate the *ex ante* value of testing at a predicted level of risk. And at a threshold of $150,000 per life year, 52.6% of tests can be flagged as wasteful before they are ever performed (Neumann, Cohen, and Weinstein 2014). The individuated risk predictions are key here: had we taken the typical approach of looking only at average yield, we would have concluded that testing overall was somewhat cost-effective, at $135,859 per life year.

[1] We consider tests performed in or shortly after emergency visits, including treadmill, radiological, and nuclear stress tests, and catheterization. Note that a patient presenting to the emergency department with new symptoms is very different from a patient with 'stable angina' presenting for routine care, for whom the effectiveness of invasive treatment has recently been questioned (e.g., Al-Lamee et al. 2018, the ISCHEMIA trial). We focus entirely on the former case, where decades of high-quality trials show large returns to early, invasive treatment (reviewed in Amsterdam et al. 2014).

[2] We use standard cost effectiveness methods from the literature (Mahoney et al. 2002; Bavry et al. 2006) and describe our procedure in detail in the Supplement.

We also find a curious fact: many patients predicted to be high-risk go untested. The tests performed on equivalently high-risk patients prove very cost-effective. However, the tested and untested are not directly comparable even when they have the same predicted risk. Physicians observe data unavailable to the algorithm (e.g., pallor of skin). So the choice to forgo testing could also suggest the untested are less risky than their tested counterparts. We illustrate the danger of ignoring such unobserved confounders by acquiring data on ECG tracings (unavailable in claims, and rarely available in electronic records). We find that the untested are systematically less risky on these typically unobserved measures.[3]

To more credibly measure under-testing—whether these untested, seemingly high-risk patients would in fact benefit from testing—-we exploit the panel nature of our data. Failure to test truly high-risk patients would result in adverse events that follow from untreated heart attack.[4] Indeed, in the 30 days after visits, untested high-risk patients experience strikingly high rates of major adverse cardiac events: 3.8% return to care, only to be diagnosed with heart attack or resulting cardiac arrest, or have an urgent treatment intervention; an additional 1.5% drop dead. These rates of adverse events suggest untreated heart attack at rates far higher than thresholds in the clinical literature, which uses the same adverse events to set widely-used decision rules and guidelines.[5] Moreover, high-risk tested patients have far lower long-term adverse event rates than their untested counterparts at the same risk levels. These facts together suggest there is also inefficient *under*-testing.

To quantify the relative importance of over- and under-testing, we simulate the effect of optimally testing only patients above a given cost-benefit threshold.[6] At $150,000 per life-year, optimal testing would cut 52.6% of current low-value tests, while adding new high-value tests equivalent to 17.9%. Under-testing, though, has bigger efficiency costs: so 42.8% of the efficiency gains come from this smaller set of added tests. Together, these results indicate physicians both over- and under-test, as suggested in earlier empirical work (Chandra and Staiger 2007; Abaluck et al. 2016).[7]

Existing policies are not well-suited to address under-use. In the simple implicit model underlying much of health policy, inefficiencies arise when patients above a certain risk threshold are tested, but incentives cause physicians to set too low a threshold. Our results suggest a problem beyond the risk threshold: how risk itself is assessed. This mis-prediction produces a different ineffi-

[3]This selection bias is perhaps unsurprising: physicians use a great deal of information that is at best imperfectly measured—even in their notes, which are sporadic and often unusable because they contain data recorded after the patient encounter. A related problem in medicine is now increasingly appreciated (measured outcomes are selectively changed by treatment, e.g., in sepsis; see Paxton, Niculescu-Mizil, and Saria 2013), but selection bias based on *unobservable* confounders remains overlooked in clinical machine learning applications (Lakkaraju et al. 2017; De-Arteaga, Dubrawski, and Chouldechova 2018; Kallus and Zhou 2018).

[4]The natural history of heart attack is well documented, since effective treatments were developed only in the 1980s. Thus we know much about the fate of untreated patients: recurrence of heart attack, arrhythmias, and death.

[5]These are drawn from studies of decision rules (e.g., TIMI: Antman et al. 2000, GRACE: Tang, Wong, and Herbison 2007, HEART: Backus et al. 2010) and subsequent validations (e.g., Sharp, Broder, and Sun 2018; Than et al. 2011; Poldervaart et al. 2017), studies of new diagnostic technologies (e.g., CT-angiography: Litt et al. 2012), or studies of treatments to reduce risk of heart attack (e.g., with statins: Ridker et al. 2008).

[6]To do so, we employ a conservative lower bound: we assume that the realized adverse events in predictably high-risk untested patients lower-bounds the under-tested population. We consider this conservative because it assumes that under-testing is concentrated in the smallest possible number of patients, all of whom would have ex ante probability 1 of an event.

[7]Chandra and Staiger 2007 find what appears to be over- or under-treatment stemming from poor choices on the part of hospitals, and the structural model of Abaluck et al. 2016 shows counterfactual outcome distributions compatible with over- and under-testing, but both leave open the potential for selection bias due to unobservables.

ciency. High-risk patients, who should be tested, are erroneously placed below the risk threshold by physicians. Policies that seek to raise the threshold will correct over-use, but might also perversely worsen under-use. We investigate this in two ways. First, we examine cross-hospital differences in testing, a common way to study inefficiency in health care. Second, we use a day-of-the-week natural experiment: testing typically requires an overnight stay, but since hospital staffing is limited on weekends, patients who come in the day before a weekend are tested less. In both cases, we find that reductions in testing happen across the board, with the marginal patient as likely to be high- or low-risk. Low-testing hospitals test *everyone* less, and doctors working on Friday and Saturday likewise test *everyone* less, irrespective of their actual risk. So incentives to reduce care may be too blunt an instrument, and exaggerate inefficiencies due to under-testing.[8]

Finally, we explore why physicians mis-predict. To do so, we build another algorithm, to predict physician choices. Contrasting it with algorithmic predictions of true risk produces two findings about physician error.[9] First, we find evidence of bounded rationality, consistent with theories that humans use simpler or more regularized models of reality (Gabaix 2014; Camerer 2018).[10] Physicians' testing decisions appear more optimal if we allow for the possibility that they might be using a simpler model of risk. Specifically, we produce several risk models at different levels of regularization. We find that the optimal regularization for out-of-sample prediction of testing decisions is far lower than the optimal level for out-of-sample risk prediction. Second, we also find evidence that physicians rely on a representativeness heuristic (Kahneman and Tversky 1972). Suppose p is probability of heart attack, and q is probability of some other cause of a patient's symptoms (and $1-p-q$ is the probability of an entirely benign cause). A recent model of representativeness suggests physicians' prediction of heart attack risk may depend not only the objective risk p, but also on $\frac{p}{q}$ (Bordalo et al. 2016): the more distinctive heart attack is as an outcome, the more representative it is. Consistent with this, we find that, holding constant true risk, patients with multiple chronic conditions are *less* likely to be tested.Of course, there might well be other behavioral explanations for this finding. For example, physicians might have a mental accounting budget of how much time, money or energy they will invest in a patient before concluding there is no problem. These findings suggest that machine learning tools might also be helpful in revealing behavioral tendencies and serve to motivate and test behavioral theories.

2 Data and Approach

2.1 Testing for Heart Attack

Heart attack is a colloquial term for acute coronary syndrome (ACS): reduction of blood flow to the heart, due to a realized or impending blockage (i.e., plaque rupture or instability, respectively)

[8]These results have analogues in patient behavior where recent work has emphasized the need for copay design to incorporate value (value-based insurance design: Chernew, Rosen, and Fendrick 2007); and to account for behavioral hazard and not simply moral hazard (Baicker, Mullainathan, and Schwartzstein 2015; Brot-Goldberg et al. 2017; Handel and Kolstad 2015).

[9]We build on a long history of research comparing clinical vs. actuarial decision making to study physician decision making (e.g., Ægisdóttir et al. 2006; Dawes, Faust, and Meehl 1989; Elstein 1999; Redelmeier et al. 2001).

[10]Limited cognitive resources, such as memory attention or computation, can result in overly simplistic—but otherwise accurate—models, because they take in only a subset of the data or variables (Simon 1955; Gabaix 2014; Sims 2003; Gabaix 2017; Mullainathan 2002a; Bordalo, Gennaioli, and Shleifer 2017).

in the coronary arteries supplying it.[11] 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 (reviewed in Amsterdam et al. 2014). 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. These large and incontrovertible treatment effects in the emergency setting, which we study here, are different from the effects of the same interventions in "stable" coronary artery disease, i.e., patients with longstanding symptoms presenting to their primary care doctors, which have been questioned in recent trials, e.g., Al-Lamee et al. 2018, or the as-yet unpublished ISCHEMIA trial.

But in order to treat heart attack, one must first diagnose it. This is easier said than done: except in the rare, most obvious cases (called "ST-elevation ACS" or STE-ACS), life-threatening blockages can have subtle symptoms, e.g., a subtle squeezing sensation in the chest, shortness of breath, or just nausea (called "non-ST-elevation ACS" or NSTE-ACS). To make matters worse, these symptoms are common in the population seeking emergency care, often the result of benign problems like acid reflux, viral infection, or a pinched nerve in the back (see Swap CJ and Nagurney JT 2005). And the simple tests done in the emergency setting (e.g., electrocardiograms, troponin testing) are often unrevealing or ambiguous.

Thus further testing is often required: "stress testing" the heart—subjecting it to an increased workload, by asking the patient to exert herself on a treadmill, or by administering a drug—or "cardiac catheterization"—an invasive procedure in which instrumentation is inserted directly into the coronary arteries to check for blockages (and, if found, to deliver the stent for treatment of the blockage). These tests have been a key part in reducing rates of missed heart attack (Nabel and Braunwald 2012), which in the 1980s and 1990s were substantial: anywhere from 2%–11% (Pope et al. 2000; Schor S et al. 1976; Lee et al. 1987) of heart attacks, meaning that patients were deprived of the benefits of timely treatment. An important technical detail to keep in mind is that stents are delivered via the same procedure as cardiac catheterization: after accessing the coronary arteries, the doctor first injects radio-opaque dye, to visualize the presence and location of the blockage, before inserting the stent at the site of the lesion. Thus the definitive test and the treatment are essentially the same procedure; the only difference is whether or not a stent is left behind to treat any blockage identified by the test.

Of course, these tests also have costs: 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 risks. Of all imaging tests, stress tests carry the single highest dose of ionizing radiation (Mettler et al. 2008), which is thought to substantially raise long-term cancer risks (Brenner et al. 2003). Exercise on a treadmill in the setting of heart attack, as required for traditional stress testing, poses a "small but definite" risk of cardiac arrest (Cobb and Weaver 1986). Likewise, the definitive test, cardiac catheterization, is

[11] Formally, ACS is defined as an acute reduction in coronary blood flow leading to insufficient oxygen and cellular damage. Despite an extensive literature on this topic (see Amsterdam et al. 2014), there are many misconceptions, even among trained physicians. For example, ACS is often erroneously equated with the results of simple tests: a laboratory study (i.e., troponin) or ECG findings (e.g., ST elevation). However, subtypes of ACS like unstable angina—ACS *without* elevated troponin or specific ECG changes—show how simplistic definitions fail to capture a complex biophysical phenomenon. The difficulties in measuring ACS precisely are why several tests are often required to diagnose it, and why decisions around high-cost testing are difficult. This motivates our current study.

an invasive procedure that also involves a large dose of ionizing radiation (Betsou et al. 1998), as well as injection of intravenous contrast material that can cause kidney failure (Rich and Crecelius 1990), and carries a risk of arterial damage (Katzenschlager et al. 1995) and even debilitating stroke (Hamon et al. 2008).

2.2 Low-Value Testing

Studies of testing typically consider, on average, whether the benefits of the test are worth the costs. Across a variety of different medical settings, these studies have documented the same key empirical fact: low average yield. Nowhere is this truer than these advanced tests for heart attack, where a growing body of research documents extremely low yield—often as low as 1%–2%—in emergency patients (Foy AJ et al. 2015; Hermann et al. 2013; Rozanski et al. 2013). This means that the vast majority of tested patients receive no tangible benefits, and are instead exposed only to costs and risks to health (complications). In light of the very low implied value of testing, health policy commentators have recently begun to advocate dramatically scaling back testing—or, in some cases, eliminating it altogether (Prasad, Cheung, and Cifu 2012; Redberg 2015). Many view the low yield of testing as only one aspect of the broader problem of "low-value health care." Despite consuming a large and ever-rising share of GDP in the US—18% of GDP with 4% annual growth (Hartman et al. 2018)—Americans' health outcomes fall short relative to other developed countries. As a result, care that provides little health benefit in light of its costs has become a central concern for policy-makers, and widely cited estimates (e.g., Committee on the Learning Health Care System in America et al. 2012) put the fraction of low-value care at one third of the $3.3 trillion in annual health care spending.

The dominant explanation for low-value care comes from economics: bad incentives. Physicians get some private benefit from doing more, in the form of extra revenue or protection from malpractice risk. As a result, while individual physicians might deliver more or less care, on average they do more than they should. This view, often referred to as "moral hazard," has been the basis for some of the most significant health policy initiatives in recent memory: a central component of the Affordable Care Act was a new Medicare payment scheme to reduce providers' incentives to deliver more care. It has also formed the basis for high-profile efforts to reduce low-value care, including the American Board of Internal Medicine's "Choosing Wisely" campaign. Of note, 62% of the items labeled as low-value were diagnostic tests (Morden et al. 2014), illustrating the centrality of testing in making the empirical case for low-value care.

2.3 A Prediction Problem

Here we frame the physician's decision to test for heart attack as a prediction problem (Kleinberg et al. 2015).[12] We begin by formalizing the way in which the doctor makes her decision. In the case

[12]Machine learning tools are particularly well suited to solving these problems, and the literature contains many recent notable examples: the prediction of specific diseases (e.g., diabetic retinopathy in Gulshan et al. 2016, sepsis in Henry et al. 2015), disease progression (e.g., Ghassemi et al. 2014; Doshi-Velez, Ge, and Kohane 2014; Attia et al. 2019; Tomašev et al. 2019), policy-relevant outcomes like hospital readmission (e.g., Rajkomar et al. 2018), or the occurrence of a variety of future diagnoses (e.g., Miotto et al. 2016). Most relevant to our own work is a series of papers predicting the output of a specific medical test, e.g., biopsies for skin (Esteva et al. 2017) or lung (Ardila et al. 2019) cancer, electroencephalography for brain death (Claassen et al. 2019), and many others. For a helpful overview of the range of problems to which machine learning is being applied in medicine, see Rajkomar, Dean, and Kohane 2019.

of testing, as with any other decision, we would want her to test patients above some threshold at which the expected benefit of the test exceeds expected cost. The costs of testing are varied but at least straightforward to quantify—all tested patients pay the direct and indirect costs of testing, and face its attendant health risks—but how to quantify the benefits? Unlike medical treatments (e.g., medications, procedures), tests per se have no direct benefit; instead, they yield information, so the value of the test is related to the extent it updates the physician's beliefs.

Simply quantifying the update, however, does not meaningfully solve the problem of measuring benefits of testing, since information is not intrinsically valuable (abstracting from any psychological benefit of 'knowing'). Rather, the concrete value of the test is that it reveals information about the *benefit of interventions*: if the patient is having a heart attack, she will benefit from treatments for heart attack; if not, no benefit is possible. So the value of testing is purely derived from the decision value it creates, in targeting interventions to patients who will benefit the most.

In other words, the presence or absence of heart attack determines the payoff from treatment, and the only way to know this is to test. But whom to test? Note that the key element in the doctor's decision to test is a prediction: given what she knows about the patient, is the risk high enough to pay the cost of testing, in order to confirm her suspicion and unlock the benefits of treatment? While the physician is ultimately in the business of allocating *treatments* for disease, to do so effectively, she must make accurate *predictions* on risk of disease (or, equivalently, accurate *diagnoses*). This is what makes testing an instance of a more general "prediction policy problem" (Kleinberg et al. 2015), requiring decision makers to synthesize complex inputs into predictions on an outcome.[13]

With this in mind, we can write the benefit of testing as follows. First, note that depending on whether an individual patient is tested, we can observe one of two counterfactual outcomes: the patient's health in the world where she is tested H_i^1 or the world where she is not H_i^0.[14] Since tests produce health via the interventions they trigger for patients who are having heart attack,[15] we can write the relationship between these two as:

$$H_i^1 = H_i^0 + T_i^+ \tau_i,$$

where T_i^+, the likelihood that the test will be positive, and τ_i is the individual's treatment effect. The expected benefit of testing is simply the difference between these two states:

$$\begin{aligned} E[H_i^1 - H_i^0 | X_i] &= E[T_i^+ \tau_i | X_i] \\ &= E[T_i^+ | X_i] E[\tau_i | X_i] + Cov(T_i^+, \tau_i | X_i). \end{aligned}$$

It is easy to see that, in the absence of covariance between T_i^+ and τ_i, the benefit of testing is monotone in likelihood of a positive test. Of course, this is not necessarily the case: in particular, patients with end-stage conditions or generally poor prognoses might have lower benefit from

[13] Naturally, in making predictions and selecting patients for testing, doctors rely on factors that are both observable and unobservable, a topic to which we return in detail below. But by contrast, the yield of the test is observable, meaning that the problem is tractable using data documented in the medical record or even insurance claims.

[14] We restrict our consideration of outcomes to the year after the visit, except in the case of cost-benefit analyses where we account for remaining life expectancy using averages derived from Peeters et al. 2002.

[15] This assumes that patients would not have received interventions in the absence of testing. Practically, stents can only be delivered via catheterization: this is the definitive test for heart attack, and is used to guide placement of the stent in the coronary arteries. So interventions physically cannot be delivered without testing to determine where in the coronary arteries they should be placed.

treatment than their risk of heart attack would otherwise indicate (e.g., because of treatment side effects, general frailty, or simply their own preferences). We address this problem by excluding from our sample anyone who might have high risk but low expected benefit of treatment. Using only data available before their emergency visits, we exclude those over 80 years of age, those with prior claims for nursing home or hospice care, and those with prior diagnosis codes indicating cancer, dementia, and other poor-prognosis conditions, following the strategy outlined in Obermeyer et al. 2017.[16]

Without these patients, we believe it is reasonable to assume $Cov(T_i^+, \tau_i|X_i) \approx 0$. In effect, we are assuming that there are no other factors that both affect average treatment effect (conditional on a positive test), and the chance that the test is positive.[17] As a result, the value of the test is monotone in the patient's ex ante risk of having a positive test. Stated another way, if we are going to do a test, we would rather do it on someone with a higher vs. lower risk of testing positive ex ante, because this person is more likely to receive and benefit from downstream life-saving interventions that she would not receive in the absence of testing. Algorithmic predictions on interventions among the tested will then form the basis for our measures of inefficiency in physician decisions: for example, the decision to test predictably low-risk patients—who go on to have a negative test—suggests errors in judgment.

2.4 Predictive Inference Strategy and Threats

We define our primary prediction target as whether or not a patient received a revascularization intervention after testing. To do so, we identify, among all tested patients, which patients ultimately receive an urgent revascularization procedure.

A key benefit of conceptualizing the value of a test in this way—by its empirical likelihood of triggering an intervention—is that it allows us to answer a specific policy question: which patients should doctors test to maximize the benefits of testing? It also lets us abstract from complex issues related to the measurement of underlying heart attack. This is useful because, while the concept of heart attack seems crisp, its empirical measurement remains difficult. Most large clinical trials and prospective studies, for example, define heart attack using an adjudication process in which a committee of clinicians judges heart attack retrospectively, by reviewing hospital records, laboratory studies, imaging, electrocardiograms, and patient narratives (e.g., Jolly et al. 2015; Chen et al. 2005).[18] Even the results of stress testing and catheterization are complex to interpret: the tests provide a readout of the heart's response to increased demand, or a picture of the dynamics of blood flow in the coronary arteries, but these complex physiological dynamics are difficult to translate into well-measured outcome variables.

[16] Another potential source of negative covariance here could relate to properties of the test itself: the medical literature documents that all tests have some risk of false positives and false negatives, i.e., $p(T_i^+|y_i = 0) > 0$ and $p(T_i^-|y_i = 1) > 0$. These outcomes are defined ex post; ex ante, they are either uncorrelated with true risk (e.g., there is some baseline likelihood of a positive test, whatever the true risk) or correlated (e.g., proportional reduction in the likelihood of a positive test that affects high-risk patients more). These latter effects would need to be quite large to change the sign of the true signal of the test and induce negative covariance between true underlying risk and likelihood of testing positive. This seems unlikely—if they did, these tests would be unlikely to find such wide clinical use—so we abstract from them in the following discussion.

[17] Many of our results, including those on under-testing, would hold if we allowed for a positive covariance: people more likely to have a positive test also have a higher conditional treatment effect.

[18] The need for this adjudication process is perhaps the best evidence that ACS is not synonymous with the results of simple tests like troponin or ECGs.

Measurement error The choice of outcome (or "label") is the central decision in algorithm development. Since algorithms can replicate and even magnify the effects of mismeasurement of the dependent variable (Mullainathan and Obermeyer 2017), it is critical to consider sources of nonrandom error.

First, in the presence of incentives to over-treat, we might mis-classify some patients whose test results were ambiguous or even negative. This would, in theory, be "priced in" to treatment effects of these interventions on mortality derived from clinical trials; but since trial and "real-world" populations may differ, we simulate a range of treatment effects at the bottom of the range of those reported in the clinical and observational literature (reviewed in Amsterdam et al. 2014; Bavry et al. 2006, and detailed fully in the Supplement). Our main point estimate is taken conservatively from the lower end of this range.

Second, if doctors are biased against some group, they might be less likely to be treated. While these biases have been demonstrated in diagnosis—e.g., 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 (e.g., Schulman et al. 1999)—these biases have not been shown or suggested *after* testing, in the decision to treat patients conditional on test results. A concrete example can help illustrate this: once a catheter is inserted in the coronary arteries, whether or not a stent is placed seems unlikely to relate to the color of the patient's skin or other sources of bias. Indeed, this is one of the major advantages of anchoring the prediction on results of testing in tested patients (provided the predictor applies to the untested).

Selective labels Predicting the outcome of testing in the tested, however, raises another key econometric challenge: when it comes time to evaluate the model's performance, we observe the outcome only in the tested. But of course, we might also want to use the model to say something about the other side of the coin: high-risk patients who were *not tested*, but who might have benefited from testing. This would answer a related policy question: which patients, when untested, suffer poor outcomes that might have been prevented with earlier testing? This is the "selective labels" problem, in which patients are non-randomly selected into measurement. Our solution to this problem takes advantage of the longitudinal nature of Medicare claims (and electronic health records) to measure potential sequelae of untreated heart attack, and identify patients who experienced poor outcomes that might indicate that they would benefit from testing. In clinical trials and cohorts, this is often defined using a basket of outcomes: subsequent diagnosis or laboratory evidence of heart attack, need for a later revascularization procedure, or cardiac arrest; in some studies, death is also considered as part of this. We describe the extent of selection bias, as well as our solution, in more detail in section 4.

2.5 Data

National Medicare claims Using a nationally representative 20% sample of Medicare claims data, we identified 20,059,154 emergency department (ED) visits over a four-and-a-half-year period from January 2009 through June 2013 (we use the last half year of 2013 as a follow-up period for included visits). We excluded non-fee-for-service patients, since we do not observe their full claims history. We also exclude those with cardiac procedures (e.g., catheterization, stenting) in the 90 days prior to ED visits, in whom testing may serve a different purpose than to diagnose new heart attack. As noted above, we exclude groups of patients whose general poor health (all observed prior to their ED visit) might mandate a different approach to testing, since they might not be

healthy enough to undergo—or want—treatments resulting from testing. We exclude those over 80 years of age; those with claims for a skilled nursing facility, to identify frail elders unable to live independently; those with poor-prognosis conditions diagnosed in the year prior (e.g., metastatic cancer, dementia, etc.); and those with a hospice claim. See Obermeyer et al. 2017 for additional details and rationale. We also exclude patients who died in the ED (i.e., a discharge code of death), and patients diagnosed with heart attack in the ED who were ultimately not tested, likely reflecting either a known diagnosis or a specific reason a test was not performed (e.g., patient preference, known prior test results). Summary statistics on demographics and concurrent medical illnesses for the final sample of 4,425,247 Medicare visits by 1,602,501 patients are shown in Table 1. The equivalent table for our electronic health record (hospital) sample, described below, is in the Supplement.

For all included visits, we identified those who had testing for heart attack, and any resulting treatments, all within 10 days of visits. This window was designed to capture both tests during the ED visit, and patients referred for urgent testing after ED visits according to current guidelines (which range from, e.g., 72 hours in Amsterdam et al. 2014 to 1-2 weeks in Brown et al. 2018). One major but under-appreciated challenge in working with claims and electronic health record data is accurate measurement of clinical tests and outcomes. A seemingly straightforward concept like "stress test" or "cardiac catheterization" is represented in a range of evolving procedure codes and test results. There is no straightforward way to capture these: for example, widely cited papers on testing for heart attack use partially non-overlapping sets of 20-30 codes to identify procedures (e.g., Sheffield et al. 2013 vs. Schwartz et al. 2014 vs. Shreibati, Baker, and Hlatky 2011). The most commonly used procedure coding system (Current Procedural Terminology, adapted with some changes for use with Medicare claims as the Healthcare Common Procedure Coding System) is modified every year. This resulted in significant changes that, in our data, led to major discontinuities in testing rates for the same hospital over time as codes and coding practices changed. To deal with this, we performed a comprehensive search of the literature as well as these coding databases. We ultimately identified 59 distinct codes for catheterization and 106 for stress test (detailed in the Supplement). Relative to those typically used in the literature, these additional codes added 11% of tests and 5% of interventions coded in our dataset.

Overall, we identified 195,287 tested visits, and 4,229,960 untested visits (which are described in more detail below). Of the tested, 124,736 had stress tests (treadmill or imaging), and 84,481 had cardiac catheterization; 13,930 had both a stress test and subsequent catheterization (the latter for definitive testing and potential stent placement). Among the tested, we identified 24,126 who received stents.[19]

Hospital electronic health records For some of our analyses, noted below, we obtain electronic health records from a large urban hospital, and re-create analyses and predictive modeling similar to that described in Medicare above. Briefly, we obtained complete data (diagnoses, procedures, laboratory studies, vital signs, ED records including complaint at visit, and electrocardiograms) on all visits to a large, urban ED over a three-year period from 2010-12, a total of 177,825 visits, and 147,953 after applying similar exclusion criteria to those described in Medicare data above.

[19]We also identified 9,700 who had a coronary artery bypass surgery (CABG) within this window. As we discuss in more detail in the Supplement, many of these CABGs appear to be semi-elective procedures routed through the ED, as opposed to patients with new symptoms requiring testing. We thus perform our main analyses without these patients; sensitivity analyses including them are substantively unchanged.

We identify 4,773 visits in which patients were tested, and 143,180 in which they were not (again, described in more detail below). Of the tested, 3,105 had stress tests and 1,668 had cardiac catheterization. Of these, 738 had revascularization procedures after the initial test.

2.6 Modeling Strategy

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). Modern claims or electronic health records, however, contain a vast set of other data, which we feed into a machine learning algorithm to predict risk. In this section, we describe these data, as well as the machine learning methods used to ensure accurate out-of-sample prediction.[20]

Input Features Raw claims data are a comprehensive record of all encounters between a patient and the health care system for which payment is exchanged (e.g., a lab test, a visit to a cardiologist for high blood pressure, a therapeutic procedure); EHR data include all encounters recorded in routine clinical systems (e.g., a laboratory study's quantitative result). To transform these transaction data into variables usable in a traditional machine learning model, we aggregate them into semantically meaningful categories (by grouping ICD-9 diagnosis codes, and ICD-9 and HCPCS procedure codes into hierarchical taxonomies defined by the Agency for Healthcare Research and Quality's Clinical Classification Software),[21] and over time (by collapsing them into discrete time periods, 0-1 months and 1-36 months prior to ED visit). When forming these features, we exclude data from the three days prior to the ED visit, to avoid any leakage of information from future claims (Kaufman et al. 2012), which can occasionally be back-dated. This results in two variables for each semantically grouped diagnosis or procedure group, describing occurrences over a recent and baseline time period. We dropped variables missing in over 99.9% of the training set, leaving 2,409 predictors X_i in the model.

Training Procedure We first randomly split the sample into a training set for model development, and a hold-out set for model validation (patients with multiple visits were assigned exclusively to either the training or hold-out set, to ensure that model results were not driven by recognizing individual patients). From the training set, we also split out a small 2.5% "ensembling set," which we use to calibrate our ensemble.

In the remaining part of the training set, we fit two machine learning methods designed to handle large sets of correlated predictors in the training data: gradient boosted trees, a linear combination of decision trees (Friedman 2001), and L1-regularized logistic regression (lasso). We train each of these to predict two outcomes: (i) $Y_i^T = 1$ on the sample where $T_i = 1$—whether the test was positive, as measured by subsequent revascularization intervention in the tested; and (ii) $Y_i^U = 1$ on the sample where $T_i = 0$—adverse cardiac events in the untested (we discuss the construction of this outcome in more detail in Section 4 on untested patients below).

[20] We are necessarily brief in our description of machine learning methods; see Mullainathan and Spiess 2017 for a more thorough overview with references.

[21] As an example, this allows us to group low-level diagnosis and procedure codes (e.g., E847: Accidents involving cable cars not running on rails) into broader clinically meaningful categories (e.g., E000-E999: External Causes Of Injury).

This procedure generates four functions (one lasso and one tree-based, for each outcome) from the training set, which are applied to generate four predictions for each observation in the 2.5% ensembling set. The final step, to find the optimal weighting of these predictions, is to predict the observed outcome of testing $Y_i^T = 1$ in the ensembling set using simple (no intercept) OLS. This weighted combination forms the ensemble model that was ultimately used to generate out-of-sample predictions in the hold-out set. It can be interpreted formally as the probability of revascularization when tested, among those tested.

Evaluation Procedure Having produced a prediction function in the training sample, we analyze its performance in the randomly sampled hold-out set of visits. We begin with 1,102,742 visits by 400,564 patients, then further restrict to the 894,166 visits by 299,325 patients under 80 years old, in addition to the exclusions noted above (regarding nursing homes, other serious illnesses, etc., to isolate a population of generally healthy patients who would have no documented reason not to benefit from testing and interventions for heart attack). All results presented below are from this independent hold-out set, to which the model was never exposed in the training process.

3 Over-testing

3.1 Model Performance

A standard measure of predictive performance is AUC, the area under the receiver operating characteristic curve (formally, $p(\hat{y}_i > \hat{y}_j | y_i = 1, y_j = 0)$. This is preferable to accuracy since our outcome is rare, and a model could achieve high accuracy simply by predicting $\hat{y}_i = 0$.) AUC for this model is 0.714 (in the hold-out set) for predicting whether a given tested patient will proceed to have a revascularization intervention in Medicare data, 0.731 in electronic health records. Logistic regression with the same variables achieves AUC of 0.672. Is this a small or a large difference? For a variety of reasons, an abstract measure like AUC can fail to capture meaningful differences in prediction. For example, it measures differences in model predictions across the entire risk distribution, when we often care most about the tails in applied prediction exercises. To illustrate this, when we take the riskiest 1% of patients in both models, we find only a 13.6% overlap—i.e., the models largely disagree on who the riskiest patients are (Table 2). So which model is right? Looking at patients for whom the models disagree, those in the top 1% of the machine learning model but not the logit model have a realized risk of 46.2%; those in the top 1% of the logit model but not the machine learning model have a realized risk of only 29.3%. This is one way to see the substantial predictive advantage that machine learning has over simpler models.

3.2 Yield of Testing among Tested Patients

Our primary interest is not in abstract metrics of algorithmic performance, however. We wish to use predictions to gain insight into physician decision making. At a coarse level, model predictions do seem to correlate with doctors' decisions on whom to test: a logit of testing T_i on model-predicted risk $\hat{y}_i$ yields a coefficient of 0.596 (standard error, SE: 0.006)—doctors are over 8.59 times more likely to test patients in the highest ventile of model-predicted risk than the lowest.

The overall correlation between risk and likelihood of testing, however, does not yield much insight into the testing margin used by doctors. To do so, we compare individualized algorithmic predictions to realized yield of testing, for each tested patient in the hold-out set. Table 3 (Column

1) reveals considerable heterogeneity in risk among patients whom doctors suspect enough to decide to test: yield is approximately monotonic in predicted risk, and the model is able to identify large groups of patients with very different risks relative to the average rate of revascularization among the tested (12.4%). For example, the lowest decile of tested patients in terms of model-predicted risk had only a 2.2% revascularization rate.

To translate yield of testing into more policy-relevant units, we draw on the substantial literature on the cost effectiveness of revascularization interventions to calculate the cost per quality adjusted life year of these tests. We can then compare these to commonly used thresholds for judging whether a given intervention is cost effective or not: $50,000 is widely cited in the UK, vs. $100,000-$150,000 in the US (Neumann, Cohen, and Weinstein 2014). We describe here the basis of our cost effectiveness calculation for the tests in our sample, modeled on the strategy of Mahoney et al. 2002, and provide a fuller accounting of individual costs, benefits, parameter choices, assumptions, and references in the Supplement.

We model the doctor's decision making process as follows. She first estimates the probability that a patient is having a heart attack, $\hat{h}_i$. If $\hat{h}_i > \frac{B_i^T}{C_i^T}$, the threshold at which benefits of test T exceed costs, she proceeds with testing.[22] If the test indicates an acute or impending blockage in the coronary arteries, the patient will proceed to stenting. In this framework, the benefits of testing B^T accrue only to those who receive treatment V as a result of the test, in terms of life years B^V (unifying both longer survival and freedom from sequelae like heart failure), i.e., $B_i^T = (B_i^V | V_i = 1)$. Patients who are tested but receive no treatment incur only costs.

$$\begin{aligned} E[C] &= C_T + p(V)C_V \\ E[B] &= p(V)B_V \\ E[B-C] &= \underbrace{p(V)[B_V - C_V]}_{\text{treated: } B-C} - \underbrace{C_T}_{\text{tested: only } C} \end{aligned}$$

Table 3 (Column 2) shows the cost effectiveness of tests in units of cost per life year. We can see that the lowest-risk patients whom doctors nonetheless choose to test are strikingly low-value: for example, the bottom 10% of tests, judged by model-predicted risk, come at a cost of $616,496 per life year. This highlights a key advantage of having individual-level machine learning predictions: we can look at the value of *marginal* tests, rather than the usual approach—looking only at the *average*. This has two advantages: first, it suggests that current estimates of the extent of low-value care are under-estimates. The average test in this population costs $135,859 per life year, which might lead us to conclude that testing as a whole is barely cost effective, or slightly ineffective relative to the commonly used thresholds of $100,000-$150,000. In the lowest-risk patients, however, doctors are performing tests with predictably lower value—by a factor of nearly 5. Second, rather than designing policies at the level of testing in general, we can identify specific marginal tests to drop, at any preferred valuation for a life year: for example, at a threshold of $150,000, we would drop 52.6% of the lowest-value tests.

Just as the average yield misses highly inefficient testing in the lowest-risk patients, it also obscures another important point: the large returns to testing the highest-risk patients. In the top

[22] For simplicity, we here refer to stress testing and catheterization together as "testing." Our tally of costs combines financial costs, both direct and indirect costs like hospitalization, with the financial and life-year costs of adverse events like peri-procedural stroke in catheterization, as noted in the Supplement.

10% of tests in terms of model-predicted risk, the cost per life year is $82,621 per life year, well within usual bounds for cost effectiveness. These high-yield patients make up a large fraction of the tested. But, strikingly, they are also well-represented in the untested pool: while our predictions were built to predict yield of testing in the tested, generating predictions for the untested identifies patients whose predicted risk levels would make them highly cost effective to test. A natural question to ask is: how often are these high-risk patients tested by doctors? Column 3 of Table 3 shows that, while doctors are more likely to test higher-risk patients, a surprising finding here is that only 10.25% of patients in the highest risk decile and 14.90% of patients in the highest percentile (defined using risk thresholds in the tested), are actually tested. This raises the possibility that, in addition to over-testing, doctors may also be under-testing.

4 Under-testing

We are not the first to raise the possibility of under-use by doctors. For example, Chandra and Staiger 2007 find what appears to be over- or under-use of interventions, stemming from poor choices across hospitals. Abaluck et al. 2016 build a structural model suggesting counterfactual outcome distributions compatible with both over- and under-testing. Others have raised the issue of under-use more broadly, in health care as a whole (Baicker, Mullainathan, and Schwartzstein 2015). Under-use is also a common finding in a substantial medical literature on diagnostic error (Kohn, Corrigan, and Donaldson 2000; Graber, Franklin, and Gordon 2005; Newman-Toker et al. 2014; Singh 2013), which points to follow-up studies of poor outcomes and malpractice claims stemming from doctors' failure to test high-risk patients.

However, any effort to study under-use must grapple with a basic econometric problem: we do not observe counterfactuals. In our setting, we do not see test results for untested patients, and do not know if they ultimately would have received interventions. In tested patients, when we see that no interventions result from the test, we can reasonably conclude what would have happened had this person not been tested: no adverse events. But in untested patients, what would have happened? Can we simply assume the model's predictions transfer from the tested to the untested (i.e., assume $E[Y_i^T|X_i, T_i = 0] = E[Y_i^T|X_i, T_i = 1]$)? This would be a strong assumption because of *private information* observed by doctors but not the statistical model. Certainly, the algorithm uses a rich set of data (X_i)—in insurance claims, all prior diagnosis and procedure codes, capturing results of prior testing; and in electronic health records, we also see complex quantitative patterns underlying laboratory studies and vital signs. But even so, the doctor has far more information available to her: the patient's appearance, the ability to question and examine the patient, the results of key diagnostic tests performed in the ED, for example the electrocardiogram (ECG) waveform, that is perhaps the most fundamental clinical tool for diagnosing heart attack. All these are at best imperfectly measured and often impossible to capture in existing datasets. As a result, the high-risk untested might reflect either under-testing, or unobserved characteristics that make them actually low yield (i.e., $E[Y_i^T|X_i, T_i = 0] \ll E[Y_i^T|X_i, T_i = 1]$). Without their test results, we cannot conclude one or the other.

A rough calculation suggests the scope of the private information problem. Consider that, among 46,714 tested patients, 5,776 (i.e., 12.4%) ultimately received prompt revascularization interventions. If the 1,053,671 untested patients had the same rate of revascularization as tested patients with the same model-predicted risk, there would be 91,319 (i.e. 8.67%) revascularization interventions in untested patients. This would in turn imply that doctors were currently diagnosing

and treating only 5.94% of all acute heart attacks in patients passing through the ED. This seems implausibly low, and suggests that more investigation is needed into the mechanisms by which doctors make use of private information in their testing decision.

4.1 The Electrocardiogram as an Unobservable We Can (Sometimes) Observe

The electrocardiogram (ECGs) is the single most important tool for diagnosis of heart attack: it is available widely and immediately, can identify obvious heart attack (STE-ACS), and provides clues to more subtle heart attack syndromes. But the ECG is not typically included in statistical models: the data are often kept in separate clinical databases from the structured EHR (and it is complex to include a waveform directly into a risk model, a topic to which we return later). But of course, if the ECG (as with any other unobserved factors) drives both doctors' decisions and the yield of testing, model predictions in the untested would be inaccurate.

To illustrate this, turn to a rich electronic health record dataset where we observe ECGs in addition to the structured data commonly used in statistical models, and inspect both the testing decision, and the downstream yield of testing. We first replicate our modeling strategy from Medicare claims, with the addition of predictors available only in electronic records (e.g., labs, vital signs), and verify similar levels of predictive accuracy (details are in the Supplement). We then identify two key ECG findings noted by the cardiologist interpreting the study (using regular expression matching): "ST elevation," a finding concerning for heart attack, and "normal ECG," which cardiologists use to denote the absence of any abnormality on the study.

Table 4 (Columns 1-3) shows first that physician testing decisions depend heavily on ECG features, conditional on our usual risk prediction. For example, in the highest bin of model-predicted risk, patients with ST-elevation are 2.9 times more likely to be tested than those with high-risk ECGs (41.7% vs 14.2%, $p < 0.001$). Conversely, those with a normal ECG are 26% less likely to be tested (11.7% vs 15.8%, $p < 0.001$). Second, these decisions correlate to true risk: yield of testing also depends on ECG features, conditional on risk prediction using structured data (Columns 4-6). Patients with ST-elevation are 2.5 times more likely to receive interventions than those with high-risk ECGs (80.0% vs 31.5%, $p < 0.001$),[23] while those with a normal ECG are 44% less likely (20.9% vs 37.4%, $p = 0.004$), all conditional on risk predictions formed without ECG data. It is also important to note is that 52.2% of patients in the highest-risk quintile of predicted risk did not even have an ECG performed. Some of these decisions may represent errors of omission; however, it is also likely to indicate that in may cases these patients had no symptoms concerning for heart attack when evaluated in the ED.

One potential problem with the approach to ECG data outlined above is that the cardiologist's interpretation of the waveform is often set down days after the visit, as she reads ECGs in large batches (this ensures reimbursement for performing the ECG—even though the formal interpretation comes far too late to be used in actual decision making). This introduces the possibility that additional information, not present in the waveform but inferred from other elements of the electronic record that accompany the ECG days later, are implicitly or explicitly incorporated into the interpretation, when the cardiologist interprets the study. So using the text of the interpretation could allow future information to "leak" into prediction of a past event (Kaufman et al. 2012).

[23] The fact that rates of revascularization are far less than 1 is yet another illustration of the complexity of identifying heart attack with a single test: medical textbooks teach that ST-elevation on ECG is synonymous with heart attack, but in our data, only 59% of these patients are found to have a suspicious blockage when catheterization is performed.

To incorporate ECG information without relying on the cardiologist's interpretation, we implement a 34-layer residual neural network (a variant of the standard convolutional neural network used for deep learning, modeled on the architecture in Rajpurkar et al. 2017) to predict the outcome of testing (yield) among the tested. The model takes as inputs both the $\hat{y}$'s from the traditional model and the raw ECG signal.[24] In the holdout set, we then compare the original predicted probability of intervention $\hat{y}_i$ to the revised predictions incorporating ECG risk information $\hat{y}'_i$. Figure 1 shows a heatmap of this comparison. This shows that risk estimates (for individuals with an ECG performed) are generally adjusted downward when ECG information is incorporated: if we calculate for each observation the likelihood that the ECG-based estimate $\hat{y}'_i$ is less than the original estimate $\hat{y}_i$, $p(\hat{y}'_i < \hat{y}_i) = 0.740$ (SE: 0.006, derived from 1000 bootstrap samples).

Of course, the ECG is just one unobservable among many that can distort conclusions from a predictive model. However, it does illustrate the scale of the unobservables problem, and points to the need for different strategies for inference in untested patients.

4.2 Clinical Outcomes in Longitudinal Data

Our solution to this problem exploits an important fact about the natural history of heart attack: undiagnosed—and thus untreated—heart attack has well-known consequences that manifest over time. These are all too well known: there is a long tradition of clinical research dating well into the modern era of medicine documenting in detail the fate of patients with untreated heart attack. This is primarily because there were no effective treatments until the early 1980s, meaning that studies of untreated heart attack were common (for example, trials comparing home vs. hospital management of heart attack, and finding no difference in outcome (Mather et al. 1976; Hill, Hampton, and Mitchell 1978). Precisely because these patients were untested (and as we show, undiagnosed), their outcomes are not masked by treatments.

In our setting, while we do not know test results in the untested, we do know their eventual outcome, thanks to the longitudinal nature of Medicare claims. This allows us to form a composite outcome, typically denoted "major adverse cardiac events," which records the known set of complications of untreated heart attack: return visits for recurrence of heart attack, need for urgent revascularization interventions, arrhythmias typically seen in the wake of heart attack, and even death.[25] This is an outcome typically tracked in clinical trials of cardiovascular interventions and observational clinical research on decision rules for doctors. These studies use an approach to creating the outcome similar to the ones we use here, and have shown good agreement with expert judgment after chart review (e.g., Wei et al. 2014).

Event data Rates of these adverse events in the 30 days after visits are shown in Figure 2. The base rate of adverse events in untested patients is 1.7%. In the highest-risk decile, however, rates are higher: 3.8% return, only to be diagnosed with heart attack or the cardiac arrest that results

[24] This consists of a 10-second ECG signal for patient i, sampled at 100 Hz to generate a vector with $t = 1000$ time steps for each of $j = 3$ channels, corresponding to three simultaneous records of the electrical depolarization of the heart measured at three different points on the chest (leads II, V1, V5), as well as the $\hat{y}_i$ from the model described above.

[25] We exclude from this outcome those patients who were not tested, but who had diagnosis codes for heart attack in the ED or elsewhere on the day of their visit; these patients were presumably known to have heart attack, but were not tested due to medical characteristics or patient preferences—an assumption we return to in our analysis of EHR data below.

from it, or to have an urgent revascularization intervention; an additional 1.5% drop dead. We can gain some insight into the doctor's decision making process by looking back to the outcome of their initial encounters: most patients with future realized adverse events were sent home from the emergency department (55.1%) instead of hospitalized, implying that doctors were unaware of their risk. Drawing on a substantial literature on diagnostic error in heart attack (e.g., Wilson et al. 2014, we form a list of common conditions like acid reflux, or symptomatic diagnoses like "chest pain" that are considered compatible with missed heart attack. Looking back at the final emergency department diagnostic codes for patients with realized adverse events, we find that a majority (60.4%) were assigned one or more of these suspicious diagnoses.

Biomarker Data Since Medicare claims data are generated for billing purposes, not to document clinical outcomes, there is the possibility of significant measurement error. Incentives to "up-code"—exaggerating the severity of the underlying medical problem in order to support increased reimbursement, which has been shown in many Medicare settings (Ibrahim et al. 2018; Geruso and Layton 2015)—could inflate rates of adverse outcomes, without basis in biological reality. So to better measure the nature of adverse events in high-risk untested patients, we return to our EHR dataset (and our standard algorithmic predictor, without ECG information), where we observe biomarker data on the severity of heart attack. If patients return to care in the 30 days after an untested visit, and a doctor decides to measure the biomarker, we can precisely measure the extent of biological damage to the heart. For these reasons, we consider this estimate a lower bound on the frequency of adverse events: patients returning to another hospital, patients who die before returning to care, and patients whose doctors decide not to test for this biomarker are not measured.[26]

Among the untested patients in our sample, biomarker-provided rates of heart attack, using measured values of cardiac troponin (i.e., cTnT, which measures death of heart muscle cells), are shown in Figure 3. In the highest-risk decile, a full 18.9% have biomarkers consistent with heart attack in the 30 days after ED visits. While small elevations in biomarkers can have various causes, over a third of these elevations are substantial (i.e., cTnT ≥ 0.1). By contrast, these outcomes are extremely rare in the lower-risk deciles.

Putting Adverse Event Rates in Context Whether measured by events or by biomarkers, these adverse event rates are large relative to benchmarks from the clinical literature. Studies of decision rules (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; 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) all set thresholds for the rate of adverse cardiac events such as the ones we measure above, which are to set guidelines for testing or treatment above a certain level of risk. This research, which is used as the basis for evaluation protocols in clinics and hospitals and underlies

[26]This source of bias is not a concern in Medicare claims, which contain the universe of encounters for a patient across all hospitals in the US. It is also less of a concern for tested patients, since those with positive tests are likely to be observed in the hospital until they recover. For this analysis, as with Medicare claims, we exclude patients whose heart attack was known to doctors on the basis of a positive troponin in the ED or a diagnosis of heart attack (4,946 of the 143,180 untested). In this setting, we can verify the assumption that these patients had a known reason for which either testing or revascularization procedures were impossible, via a hand-review of a sample of charts, including patient or family preference, known severe coronary disease refractory to treatment, and other reasons.

recommendations from professional societies, can be compared to the rates we measure—5.3% with adverse events, 6.8% with substantial biomarker elevation—to gauge whether the highest-risk patients have predictable event rates high enough to mandate further evaluation in clinical practice. While studies vary in the follow-up period they consider (typically between 30 and 60 days after visits; we conservatively use 30), the thresholds that mandate action (testing, treatment, admission to the hospital, etc.) are typically under 2%. Likewise, surveys of practicing emergency doctors have suggested that they would tolerate at most a miss rate of 1% (Than et al. 2013). Since the rates we observe in high-risk patients are higher than these thresholds, we believe that testing these high-risk untested patients would be seen as cost-effective by the standards widely used in clinical research and practice.

With our data, we can go further. These guidelines are based on the assumption that the benefits of testing are proportional to a patient's underlying risk (as measured by these adverse event rates in undiagnosed and untreated patients). Using our risk predictions, we can check this assumption directly. Analagous to a propensity score, we can directly compare long-term adverse rates between tested and untested stratified by predicted risk. To make this comparison, we first define a follow-up window for longer-term adverse events, from 31 to 365 days after the visit.[27] We then calculate the rates of one-year MACE in the tested and untested groups, by predicted risk.

The assumption that the benefits of testing and treatment are monotonic in the risk of adverse events has a clear prediction. Among low-risk patient pools, testing should not matter: since tests are unlikely to lead to diagnosis and treatment, we should detect no differences in longer-term outcomes between tested and untested patients. But as risk of heart attack increases, we should observe a clear and increasing return to testing, in terms of reduced rates of later adverse events. Figure 4 shows in fact no significant difference in adverse event rates for patients in deciles up to the seventh. However, starting with the eighth risk decile, there is an increasingly large difference in MACE rate, until the highest-risk decile, where tested patients have a 5.72-percentage-point lower adverse event rate than untested patients (SE: 0.93 percentage points)—a relative difference of 43.6%. Of course, there are likely unobserved differences between tested and untested groups above and beyond risk. But as we show above, physician private information works against finding an effect here: in general, as we show above, tested patients are unobservably *higher* risk than untested patients, and yet tested patients despite this have better long-term outcomes. This illustrates both the extent of low-value testing, and the returns to testing high-risk patients.

4.3 Over- and Under-testing

While we have found evidence of both over- and under-testing, we would ideally like to compare their magnitude. Over-testing is straightforward to measure, by applying standard cost effectiveness calculations and thresholds to tested patients. Under-testing is more difficult: how to translate adverse event rates into a quantitative estimate of under-testing? To do so, we estimate a simple lower bound. Observe that the lowest rates of under-testing would be if all adverse outcomes were concentrated (with $p = 1$) in an ex ante identifiable set of people. In other words, a conservative estimate of under-testing is to consider only those untested high-risk patients who go on to have

[27]The clinical guidelines referenced above are meant to detect short-term risk, as manifested in adverse events over the 30 days after emergency visits. Here, by contrast, we are interested in the effects of testing and the resulting treatments that may accrue over a longer period since the visit. We omit the first month to avoid coding the interventions and follow-up visits in tested patients as adverse events, when they were in fact the intended results of testing.

realized adverse events, as those who would have been likely to benefit from testing. To estimate the cost effectiveness of testing them, we simply use the cost effectiveness implied by their ex ante risk (estimated in the tested).

We can then calculate, at different thresholds for cost effectiveness, the net amount of over- and under-testing, as shown in the first panel of Figure 5. For example, at a cost per life year valuation of $150,000, there is both substantial over- and under-testing: we would drop the 52.6% of tests doctors currently do, but we would also add back 17.9% (relative to the current number of tests) for high risk patients not currently tested. The second panel of Figure 5 shows the results of the same procedure, but translated into units of dollar benefits rather than number of tests. This combines the number of life years saved (at the dollar valuation on the x-axis) and the costs of testing. Importantly, even though this strategy would on net *reduce* testing, a large fraction of the benefits of this reallocation comes from *increasing* testing for the high-risk untested. For example, at $150,000 per life year, we would reduce testing by 34.7% on net—but 42.8% of welfare gains come from remedying under-testing (i.e., $228.0 million in surplus from life years saved), as opposed to reducing over-testing (i.e., $304.7 million saved from dropping low-value tests). This fraction of benefits from increasing testing grows with the valuation of a life year.

4.4 Implications for Policy: Hospital Variation

The simultaneous existence of both over- and under-testing highlights the deficiencies of a model built solely on incentives. In models of physician or health provider moral hazard, doctors test patient above a threshold level of risk. Because of incentives to test, they choose a threshold that is too low; the primary inefficiency then is that many low-risk patients are tested. Such a model has a clear remedy: adjust incentives or implement other policies to raise the threshold physicians use. But such a model cannot explain our finding of a failure to test high-risk patients.

Allocating tests to low-risk patients instead of high-risk patients is particularly hard to explain from a pure revenue-maximizing perspective: patients with higher risk of heart attack are more likely to generate the complex procedures and intensive care needs for coronary care that are major contributors to hospitals' bottom lines Abelson and Creswell 2012, and thus more profitable than a negative test. As a result, policies built on this model may have perverse consequences in a world where over- and under-testing co-exist.

We test this by replicating a standard analysis in health economics, cross-hospital variation, but this time through the lens of algorithmically predicted risk. A long tradition of research has documented wide variations in the amount of health care delivered, but little variation in outcomes. The most commonly cited explanation for this is that marginal patients are low-value, so testing them will generate more care, but not better outcomes. And the most commonly cited solution is policies that change provider incentives (e.g., Obamacare): the idea is to encourage doctors to test less, but let them decide where and when to cut back care (e.g., Loewenstein, Volpp, and Asch 2012). The implicit goal is for high-utilizing providers to look like their low-utilizing neighbors (e.g., Liao, Fleisher, and Navathe 2016).

However, as shown in Figure 6, high- and low-utilizing hospitals are simply inefficient in different ways. The top panel of the Figure first groups all US hospitals into quintiles, based on their empirical testing rate in our sample, then shows how testing rates vary across hospitals by bin of predicted risk. Strikingly, when doctors test more (or less), they test *everyone* more (or less). Marginal patients are drawn from across the entire risk distribution, not just low-risk groups. The bottom panel shows a similar pattern using variation linked to hospital ownership: teaching

hospitals test everyone more, federal hospitals (e.g., Veterans Affairs, Indian Health Services) test everyone less, and for-profit and non-profit hospitals are somewhere in the middle.

4.5 Day-of-Week Differences

This evidence suggests that the "marginal" patient—the kind we would reduce testing of when policies reduce testing—may come across from the risk distribution. Of course, since other factors may vary between hospitals, it does not necessarily identify the marginal patient. While it is common in health policy research to compare variation in care delivered by providers——doctors, hospitals, regions, etc.——these comparisons can be challenging, because patients may differ on unobservable characteristics: even after adjustment for a variety of factors (e.g., our own risk predictions), we cannot be sure that there remain other unaccounted-for differences.

To more precisely study which patients are "marginal," we take advantage of a natural experiment in testing for heart attack that takes place in hospitals across the US every weekend. The typical testing protocol for emergency patients with suspected heart attack is to observe them overnight, then perform definitive testing the following morning. This is both because it takes time to arrange the test, and because of the need to observe patients for stability: there is an elevated risk of sudden death if tests are done on unstable patients, so patients are typically monitored overnight, undergo repeat laboratory testing (troponin), then have their test done on the next day. However, because it is expensive to maintain staffing of cardiac testing facilities, many hospitals leave them unstaffed on weekends (Krasuski et al. 1999). While testing is still available if the doctor on duty makes the decision to call in the team from home, it is widely assumed to require a higher threshold for doing so.

As a result, we hypothesized that patients who come in on the day *before* a weekend day—i.e., Friday or Saturday—would be less likely to be tested. This strategy builds on prior research showing differences in care for patients admitted on weekends vs weekdays (Bell and Redelmeier 2001), but has the additional advantage of straddling a weekend (Saturday) and weekday (Friday), which reduces the risk of confounding by simply comparing weekend patients to weekday patients. To reduce other sources of bias, we also wished to exclude patients who had been transferred or referred to specialized hospitals from other facilities, for whom decision making might be less sensitive to inconveniences related to in-hospital staffing. Practically, we restricted our sample to hospitals with a catheterization laboratory on-site (using the American Hospital Association annual survey data), and to patients whose home zip codes are within 10 miles of these facilities, to zero in on patients presenting to hospitals near their home zip code that had on-site testing facilities. In this sample, we find that patients are 19.8% less likely to be tested when their index visit falls on Fridays and Saturdays than on Sundays through Thursdays (3.95% vs. 4.93%, $p < 0.001$). Figure 7 shows that, conditional on geography (i.e., hospital referral region) and year, these patients appear otherwise quite similar on observables. There are small differences in some risk factors for heart disease: while some of these are statistically significant after Bonferroni adjustment (the unadjusted Figure is in the Supplement), they are substantively small, most on the order of < 0.01 SD units and statistically insignificant. Finally, as a summary statistic, there is only a very small (0.01 SD) difference in overall risk, measured by $\hat{y}$, that is also statistically insignificant, meaning that many small differences in individual variables largely balance out.

We first verify that the model predicts accurately in this setting: rates of realized yield in the tested and adverse events in the untested are similar in both pre-weekend and pre-weekday

patients. This is shown in Figure 8.[28] We also verify the relationship between yield (among the tested) and adverse event rates (among the untested), in each percentile bin of $\hat{y}$: this is monotonic and approximately linear in this weekend vs. weekday population. This gives us some suggestive evidence that we are measuring the same underlying latent risk, manifested differently depending on whether doctors decide to test or not (Figure 9).

Finally, in this setting with limited influence of unobservables, we can more precisely answer the question: when doctors reduce testing by 19.8%, where in the risk distribution do the marginal patients come from? The results in Figure 10 echo the results from our (less well-identified) cross-sectional analysis of hospital differences. We see again that when doctors reduce testing on weekends, they drop marginal patients from across the risk distribution, not just low-risk patients. For example, patients in the lowest-risk decile are 14.3% less likely to be tested on a weekend (2.2% vs 2.6%, $p < 0.001$); patients in the highest-risk decile are 22% less likely to be tested (10.0% vs 12.8%, $p < 0.001$). This suggests that, when doctors cut back on testing, they do so fairly indiscriminately. The Figure also shows, by comparison, how much better an algorithm would do in allocating testing in this setting, where we believe the influence of unobservables to be minimal. Specifically, we identify the lowest-risk patients seen on weekdays within a geography-time bin (within which patients are observably and hopefully unobservably similar), and drop these until we get to a 19.8% reduction in testing. Our findings suggest substantial scope for improvement: relative to doctors, the algorithm would drop 245.5% more of the lowest-risk patients (lowest decile: 35.5 vs 14.4%, $p < 0.001$), and drop 90.0% fewer of the highest-risk patients (highest decile: 1.4% vs 13.7%, $p < 0.001$). In other units, to achieve the same decrement in testing, the algorithm would drop tests with a mean cost effectiveness of \$281,720, vs doctors, whose marginal tests have a cost effectiveness of \$198,964. Since we observe the test results in the dropped patients, we can confidently translate these policy simulations into counterfactual comparisons: at the same level of reductions in testing, an algorithmically driven strategy would find 17.3% more patients needing intervention than doctors, saving 17.7% more life years and generating (at \$150,000 per life year) a surplus of \$150.7 million in our sample.

5 Cognitive Errors

Our results highlight the need for new ways to understand the mechanisms that lead doctors to mis-predict, which seem incompatible with a model based on incentives alone. Machine learning can help us shed light on this question as well. Specifically, we use the contrast between algorithmic predictions, trained to predict true risk (as captured by the outcomes of testing), and physician judgment (as revealed by their testing decisions), to investigate potential mechanisms for error. In doing so, we build on a long tradition of research comparing clinical judgment to statistical models as a way to gain insights into physician decision making (Ægisdóttir et al. 2006; Dawes, Faust, and Meehl 1989; Elstein 1999; Redelmeier et al. 2001.

[28] We attempted to perform a similar analysis on marginal yield and adverse event rates, by subtracting out from weekday and weekend rates, respectively, the fraction we would expect based on weekend and weekday rates, respectively. These results unfortunately suffer from large imprecision due to small samples, after sample restrictions, separation into bins of $\hat{y}$, and further restriction to marginal patients.

5.1 Boundedness

To put our findings in context, we differentiate between two categories of explanations for errors in judgment. The first, going back to at least Herbert Simon (Simon 1955), points to the boundedness of human cognition. For example, people may have bounded memory Mullainathan 2002a; Bordalo, Gennaioli, and Shleifer 2017, use coarse categories rather than specific models (Rosch 1999; Mullainathan 2002b; Mullainathan, Schwartzstein, and Shleifer 2008), or have limited attention that constrains their ability to code all the variables in their environment (Gabaix 2017; Gabaix 2014; Taubinsky and Rees-Jones 2018; Chetty, Looney, and Kroft 2009; Sims 2003). Boundedness is very natural in our setting: physicians may not be able to attend to, process, or mentally represent the rich set of data available on their patients, and so may instead resort to a simpler model of risk. By contrast, a second set of models emphasizes errors: even within the set of variables people do use in their mental models, they make systematic mistakes. For example, even when all variables are attended to and well represented, people are known to overweight certain ones or even make basic errors in probability (Tversky and Kahneman 1974).

The machine learning toolbox provides one way to test these two models of human judgment. Our basic insight is to create a series of progressively simpler models of true risk in the training set, using regularization: specifically, we use a LASSO to predict intervention among the tested, as above, and preserve all models along the regularization path, from OLS estimates (i.e., no regularization, $\lambda = 0$) to constant prediction ($\lambda \to \infty$). This set represents models with a range of complexity, here parameterized by the sum contributions of all variables to predictions (the target for regularization), ranging from 2,093 variables (OLS estimates with the full set of X's) to 0 variables (simply predicting the base rate for everyone).

We first quantify how well these models predict the outcome in our holdout set, as shown in Figure 11 and measured by the area under the curve (AUC, again chosen to measure predictive performance for a rare outcome). As we expect, the accuracy of predictions increases in the number of included variables until 299 variables, the best-performing model (accuracy then decreases as the remaining variables begin to induce over-fitting). We then assess how well these same models predict another outcome: not the ground truth, but the doctor's testing decision. Here, a different picture emerges. The model of ground truth that best predicts the doctor's decision is not the best-performing one (299 variables), but rather a highly simplified model of true risk with only 21 variables. Thus doctors appear to use a far simpler model of risk than the algorithm.

Importantly, the variables included in this simpler model are weighted proportionally to their observed relationship to true risk—in other words, doctors appear generally to get the signs and magnitudes of these variables right. This is not necessarily the case: doctors could be using any number of variables that are negatively correlated with the ones incorporated into the simple model, meaning high predictability but very different weightings. But in fact, as shown in Figure 12, the correlation between the doctor's weights and the algorithm's is 0.812.

These results provide some support for boundedness in physician judgments. Doctors, much like algorithms, appear to be regularizing (Camerer 2018; Gabaix 2014): they are identifying a small number of good risk predictors and using them if not perfectly, at least quite well. But conversely, they are neglecting 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 emphasizes simple statistical models; whereas we find

that extremely complex, high-dimensional models provide a major advantage and in predictive performance.

5.2 Representativeness

Though these results provide evidence that physicians are using a small number of variables effectively, this does not imply the absence of specific biases. One such bias we explore is related to the idea of representativeness (Tversky and Kahneman 1974), as formalized in the model of stereotyping of Bordalo et al. 2016.

Imagine that there are two types of patients presenting to the emergency department, each with a new symptom like chest pain. Patient s is generally healthy, while patient c has several chronic illnesses. Objectively define the patients' probability of heart attack given the chest pain as p_s and p_c. Of course, chest pain can have many causes, so let q_s and q_c be the probability that the pain indicates some other clinically important problem (and $1 - q - p$ is the probability that there is no problem at all). Since the complex patient has more illnesses, we assume that $q_c > q_s$. Applying the Bordalo et al. 2016 model of stereotypes here, we would assert that doctors form probability judgments using $p_s h(\frac{p_s}{q_s})$ and $p_c h(\frac{p_c}{q_c})$, where $h()$ is a monotonic function. So holding constant true risk of heart attack, $p_s = p_c$, the judged probability also depends on the distinctiveness of heart attack relative to other possible causes. Since $q_c > q_s$, this model has a crisp empirical prediction: at the same true risk, patients with other potential conditions are *less* likely to be tested.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 there are many other conditions that might explain a patient's symptoms, physicians may be tempted to "close early" on one, thereby failing to consider another.

Figure 13 shows that this hypothesis fits the data we observe. We first restrict the sample to the 44.2% of patients who had not been previously diagnosed with heart attack or stroke, to identify those in whom a new diagnosis is likely to have the greatest implications for treatment (we show in the Supplement that these patterns hold similarly in the full population). We then graph testing rate vs. predicted risk, for three approximately equal-sized bins measuring number of chronic conditions (using a standard approach in the literature, outlined in Gagne et al. 2011; we graph predicted testing rate, based on patient observables and predicted using the machine learning approach described above, rather than actual testing rate, since the latter was visually noisy in the smaller comorbidity-risk bins.)

In Table 5, we show the regression analogues of this figure. We regress the test decision on algorithmically predicted risk and a set of other variables capturing the extent and nature of other illnesses (here we use actual testing decisions, rather than the predicted probabilities shown in the graph). Column 1 shows that the likelihood of testing decreases by 0.1 percentage point (SE: 0.0002) with each additional illness (the mean testing rate is 6.1%). Column 2 divides patients into three roughly equal-sized bins, based on the number of illnesses, indicating that those in the highest bin (9 or more illnesses) are 1.4 percentage points less likely to be tested (SE: 0.002). To scale this difference in testing rates, consider that for patients in the highest decile of risk, complex patients (in the top third of the chronic condition distribution) are tested at the same rate as simpler patients in the seventh risk decile—i.e., patients who have approximately 30% lower objective risk.

Column 3 explores the differential effect of two types of chronic illnesses: those that are directly related to cardiovascular risk (e.g., high blood pressure, cholesterol, heart failure) *increase* testing

conditional on risk by 0.6 percentage points each (SE: 0.001), while non-cardiovascular illnesses (e.g., pulmonary diseases, depression, alcohol abuse) decrease testing by 0.4 percentage points (SE: 0.0003). The Supplement provides additional information on associations of individual conditions with testing rates conditional on risk.

Finally, columns 4-6 show the same analyses, but with the yield of testing on the left-hand side, showing that differences in chronic conditions have no significant association with the yield of testing when these patients are tested. This indicates that the differences in testing rate are unlikely to be a function of physician private information.

6 Conclusions

Much of our understanding of the health care system has its roots in how we model physician behavior. There is increasing evidence that our current models cannot explain the widespread inefficiencies observed in patients' decision making (Baicker, Mullainathan, and Schwartzstein 2015; Brot-Goldberg et al. 2017; Handel and Kolstad 2015). 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).

This has implications for how interventions can be devised to reduce low-value care. To date, policy makers have aimed squarely at sources of moral hazard, largely around the incentives of providers; often, these interventions simply reduce global rates of reimbursement for services. 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 (Newhouse and Group 1993) and more recent work since (Brot-Goldberg et al. 2017) has shown, changing incentives cuts all care, not just low-value care. If similar problems affect interventions aimed at physicians, these policies may have less impact than hoped. The ability to predict the value of a specific medical intervention for a specific person 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 et al. 2018): comparing idealized predictions to the actions of individual actors is a fascinating new lens through which to view human behavior in complex environments.

References

Abaluck, Jason et al. (2016). "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care". In: *American Economic Review* 106.12, pp. 3730–3764.

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

Al-Lamee, Rasha et al. (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. et al. (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. et al. (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.

Ardila, Diego et al. (June 2019). "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography". In: *Nature Medicine* 25.6, pp. 954–961.

Attia, Zachi I. et al. (Sept. 2019). "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction". In: *Lancet (London, England)* 394.10201, pp. 861–867.

Backus, Barbra E. et al. (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.

Bavry, Anthony A. et al. (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.

Bell, Chaim M. and Donald A. Redelmeier (Aug. 2001). "Mortality among Patients Admitted to Hospitals on Weekends as Compared with Weekdays". In: *New England Journal of Medicine* 345.9, pp. 663–668.

Betsou, S. et al. (1998). "Patient radiation doses during cardiac catheterization procedures." In: *The British journal of radiology* 71.846, pp. 634–639.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2017). "Memory, attention, and choice". In: *National Bureau of Economic Research Working Paper*.

Bordalo, Pedro et al. (2016). "Stereotypes". In: *The Quarterly Journal of Economics* 131.4, pp. 1753–1794.

Brenner, David J. et al. (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. et al. (2017). "What does a deductible do? the impact of cost-sharing on health care prices, quantities, and spending dynamics". In: *The Quarterly Journal of Economics* 132.3, pp. 1261–1318.

Brown, Michael D. et al. (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.

Chandra, Amitabh and Douglas O. Staiger (Feb. 2007). "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks". In: *Journal of Political Economy* 115.1, pp. 103–140.

Chen, Z. M. et al. (Nov. 2005). "Addition of clopidogrel to aspirin in 45,852 patients with acute myocardial infarction: randomised placebo-controlled trial". In: *Lancet (London, England)* 366.9497, pp. 1607–1621.

Chernew, Michael E., Allison B. Rosen, and A. Mark Fendrick (2007). "Value-Based Insurance Design". In: *Health Affairs* 26.5, w195–w203.

Chetty, Raj, Adam Looney, and Kory Kroft (2009). "Salience and taxation: Theory and evidence". In: *American economic review* 99.4, pp. 1145–1177.

Claassen, Jan et al. (June 2019). "Detection of Brain Activation in Unresponsive Patients with Acute Brain Injury". In: *New England Journal of Medicine* 380.26, pp. 2497–2505.

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.

Committee on the Learning Health Care System in America et al. (2012). *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Washington: National Academies Press.

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.

De-Arteaga, Maria, Artur Dubrawski, and Alexandra Chouldechova (2018). "Learning under selective labels in the presence of expert consistency". In: *arXiv preprint arXiv:1807.00905*.

Doshi-Velez, Finale, Yaorong Ge, and Isaac Kohane (Jan. 2014). "Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis". In: *Pediatrics* 133.1, e54–e63.

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

Esteva, Andre et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639, pp. 115–118.

Foy AJ et al. (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.

Gagne, Joshua J et al. (July 2011). "A combined comorbidity score predicted mortality in elderly patients better than existing scores". In: *Journal of clinical epidemiology* 64.7, pp. 749–759.

Geruso, Michael and Timothy Layton (2015). "Upcoding: Evidence from Medicare on squishy risk adjustment". In: *National Bureau of Economic Research Working Paper*.

Ghassemi, Marzyeh et al. (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.

Gulshan, Varun et al. (2016). "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs". In: *JAMA* 316.22, pp. 2402–2410.

Hamon, Martial et al. (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.

Hartman, Micah et al. (Dec. 2018). "National Health Care Spending In 2016: Spending And Enrollment Growth Slow After Initial Coverage Expansions". In: *Health Affairs* 37.1, pp. 150–160.

Henry, Katharine E. et al. (2015). "A targeted real-time early warning score (TREWScore) for septic shock". In: *Science translational medicine* 7.299, 299ra122299ra122.

Hermann, Luke K. et al. (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)* 311.8069, pp. 837–841.

Ibrahim, Andrew M. et al. (Feb. 2018). "Association of Coded Severity With Readmission Reduction After the Hospital Readmissions Reduction Program". In: *JAMA Internal Medicine* 178.2, pp. 290–292.

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

Jolly, Sanjit S. et al. (Apr. 2015). "Randomized Trial of Primary PCI with or without Routine Manual Thrombectomy". In: *New England Journal of Medicine* 372.15, pp. 1389–1398.

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

Kallus, Nathan and Angela Zhou (2018). "Confounding-Robust Policy Improvement". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., pp. 9269–9279.

Katzenschlager, Reinhold et al. (1995). "Incidence of pseudoaneurysm after diagnostic and therapeutic angiography." In: *Radiology* 195.2, pp. 463–466.

Kaufman, Shachar et al. (2012). "Leakage in data mining: Formulation, detection, and avoidance". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6.4, pp. 1–21.

Kleinberg, Jon et al. (2015). "Prediction Policy Problems". In: *American Economic Review* 105.5, pp. 491–95.

Kleinberg, Jon et al. (2018). "Human decisions and machine predictions". In: *The Quarterly Journal of Economics* 133.1, pp. 237–293.

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.

Krasuski, Richard A et al. (Jan. 1999). "Weekend and Holiday Exercise Testing in Patients with Chest Pain". In: *Journal of General Internal Medicine* 14.1, pp. 10–14.

Lakkaraju, Himabindu et al. (2017). "The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 275–284.

Lee, Thomas H. et al. (1987). "Clinical characteristics and natural history of patients with acute myocardial infarction sent home from the emergency room". In: *The 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. et al. (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. et al. (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 et al. (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. et al. (July 2008). "Effective doses in radiology and diagnostic nuclear medicine: a catalog". In: *Radiology* 248.1, pp. 254–263.

Miotto, Riccardo et al. (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. et al. (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 (2002a). "A memory-based model of bounded rationality". In: *The Quarterly Journal of Economics* 117.3, pp. 735–774.

— (2002b). "Thinking through categories". In: *National Bureau of Economic Research Working Paper*.

Mullainathan, Sendhil and Ziad Obermeyer (2017). "Does Machine Learning Automate Moral Hazard and Error?" In: *American Economic Review* 107.5, pp. 476–480.

Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer (2008). "Coarse thinking and persuasion". In: *The Quarterly journal of economics* 123.2, pp. 577–619.

Mullainathan, Sendhil and Jann Spiess (2017). "Machine Learning: An Applied Econometric Approach". In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Nabel, Elizabeth G. and Eugene Braunwald (Jan. 2012). "A Tale of Coronary Artery Disease and Myocardial Infarction". In: *New England Journal of Medicine* 366.1, pp. 54–63.

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. et al. (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 et al. (Feb. 2017). "Early death after discharge from emergency departments: analysis of national US insurance claims data". In: *BMJ* 356, j239.

Papanicolas, Irene, Liana R. Woskie, and Ashish K. Jha (Mar. 2018). "Health Care Spending in the United States and Other High-Income Countries". In: *JAMA* 319.10, pp. 1024–1039.

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, pp. 1109–1115.

Peeters, Anna et al. (2002). "A cardiovascular life history: : A life course analysis of the original Framingham Heart Study cohort". In: *European heart journal* 23.6, pp. 458–466.

Poldervaart, J. M. et al. (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 et al. (2000). "Missed diagnoses of acute cardiac ischemia in the emergency department". In: *New England Journal of Medicine* 342.16, pp. 1163–1170.

Prasad, Vinay, Michael Cheung, and Adam Cifu (2012). "Chest pain in the emergency department: : the case against our current practice of routine noninvasive testing". In: *Archives of Internal Medicine* 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 et al. (2018). "Scalable and accurate deep learning with electronic health records". In: *NPJ Digital Medicine* 1.1, p. 18.

Rajpurkar, Pranav et al. (2017). "Cardiologist-level arrhythmia detection with convolutional neural networks". In: *arXiv preprint arXiv:1707.01836*.

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, p. 436.

Redelmeier, Donald A. et al. (Feb. 2001). "Problems for clinical judgement: introducing cognitive psychology as one more basic science". In: *CMAJ: Canadian Medical Association Journal* 164.3, pp. 358–364.

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 et al. (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.

Rosch, Eleanor (1999). "Concepts: Core Readings". In: ed. by Stephen Laurence Eric Margolis. MIT Press. Chap. Principles of categorization, pp. 189–206.

Rozanski, Alan et al. (Mar. 2013). "Temporal Trends in the Frequency of Inducible Myocardial Ischemia During Cardiac Stress Testing: 1991 to 2009". In: *Journal of the American College of Cardiology* 61.10, pp. 1054–1065.

Schor S et al. (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. et al. (1999). "The effect of race and sex on physicians' recommendations for cardiac catheterization." In: *The New England journal of medicine* 340.8, pp. 618–626.

Schwartz, Aaron L. et al. (July 2014). "Measuring Low-Value Care in Medicare". In: *JAMA Internal Medicine* 174.7, pp. 1067–1076.

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

Sheffield, Kristin M. et al. (2013). "Overuse of preoperative cardiac stress testing in medicare patients undergoing elective noncardiac surgery". In: *Annals of surgery* 257.1, pp. 73–80.

Shreibati, Jacqueline Baras, Laurence C. Baker, and Mark A. Hlatky (Nov. 2011). "Association of Coronary CT Angiography or Stress Testing With Subsequent Utilization and Spending Among Medicare Beneficiaries". In: *JAMA* 306.19, pp. 2128–2136.

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.

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.

Taubinsky, Dmitry and Alex Rees-Jones (2018). "Attention variation and welfare: theory and evidence from a tax salience experiment". In: *The Review of Economic Studies* 85.4, pp. 2462–2496.

Than, Martin et al. (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 et al. (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.

Tomašev, Nenad et al. (Aug. 2019). "A clinically applicable approach to continuous prediction of future acute kidney injury". In: *Nature* 572.7767, pp. 116–119.

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

Wei, Wei-Qi et al. (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.

Wilson, Michael et al. (Oct. 2014). "Hospital and Emergency Department Factors Associated With Variations in Missed Diagnosis and Costs for Patients Age 65 Years and Older With Acute Myocardial Infarction Who Present to Emergency Departments". In: *Academic Emergency Medicine* 21.10, pp. 1101–1108.

Ægisdóttir, Stefanía et al. (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

Variable	All	Tested	Untested
n Patients	1,556,477	150,616	1,508,267
n Visits	4,246,642	189,290	4,057,352
Demographics			
Age, mean	63	68	63
Age, median [IQR]	66 [49,77]	70 [60,77]	66 [49,77]
Female (%)	0.593	0.554	0.594
White (%)	0.763	0.786	0.762
Black (%)	0.181	0.162	0.181
Hispanic (%)	0.028	0.023	0.028
Other (%)	0.029	0.030	0.028
Distance to hospital, median [IQR]	7 [2,16]	8 [3,17]	7 [2,16]
Eligibility			
Aged in	0.440	0.555	0.435
Disability	0.541	0.426	0.547
Risk factors			
Atherosclerosis (%)	0.562	0.723	0.556
Cholesterol (%)	0.662	0.800	0.656
Diabetes (%)	0.485	0.555	0.482
Hypertension (%)	0.813	0.901	0.810

Table 1: Medicare sample descriptive statistics.

	Overlap %	Yield % (SE)				
Risk Percentile	ML ∩ logit (1)	ML ∩ logit (2)	ML only (3)	Logit only (4)	ML all (5)	Logit all (6)
Top 1	50.7	50.1 (2.54)	48.3 (2.57)	32.9 (2.42)	49.2 (1.81)	41.6 (1.78)
Top 5	61.8	42.2 (1.02)	36.7 (1.26)	28.1 (1.18)	40.1 (0.793)	36.8 (0.780)
Top 10	65.9	37.6 (0.683)	31.9 (0.913)	25.7 (0.856)	35.6 (0.548)	33.5 (0.540)
Top 25	73.7	31.2 (0.391)	24.7 (0.609)	21.1 (0.576)	29.5 (0.330)	28.6 (0.327)
Bottom 1	45.4	8.4 (1.49)	6.7 (1.23)	6.2 (1.18)	7.5 (0.951)	7.2 (0.935)
Bottom 5	52.0	6.9 (0.567)	6.4 (0.571)	9.7 (0.691)	6.6 (0.402)	8.2 (0.445)
Bottom 10	57.8	6.7 (0.376)	6.1 (0.422)	10.4 (0.539)	6.4 (0.281)	8.3 (0.315)
Bottom 25	68.9	7.2 (0.226)	9.0 (0.371)	13.2 (0.439)	7.8 (0.194)	9.1 (0.208)

Table 2: Comparison of machine learning (ML) estimates vs. logit estimates fit with same vector of predictors. For each risk group (rows), we quantify the overlap—the degree to which ML and logit models agree regarding which patients are in the group. We then quantify the realized yield for patients in the intersection of both, patients only in the ML vs logit high (low) risk groups, and realized yield in all patients in the ML vs logit groups.

Risk Ventile	Yield (SE) (1)		Cost, $ (2)	Test rate (SE) (3)	
1	0.017	(0.003)	650,838	0.015	(0.000)
2	0.022	(0.003)	587,572	0.024	(0.000)
3	0.034	(0.004)	366,289	0.030	(0.001)
4	0.049	(0.005)	270,292	0.036	(0.001)
5	0.063	(0.005)	222,940	0.042	(0.001)
6	0.082	(0.006)	178,145	0.043	(0.001)
7	0.075	(0.006)	181,552	0.048	(0.001)
8	0.076	(0.006)	203,132	0.048	(0.001)
9	0.092	(0.007)	165,491	0.052	(0.001)
10	0.094	(0.007)	171,460	0.053	(0.001)
11	0.114	(0.007)	140,606	0.056	(0.001)
12	0.124	(0.008)	140,064	0.061	(0.001)
13	0.145	(0.008)	119,263	0.064	(0.001)
14	0.143	(0.008)	131,469	0.064	(0.001)
15	0.158	(0.008)	121,253	0.070	(0.002)
16	0.193	(0.009)	105,463	0.075	(0.002)
17	0.199	(0.009)	102,103	0.079	(0.002)
18	0.206	(0.009)	103,568	0.089	(0.002)
19	0.254	(0.010)	90,504	0.103	(0.002)
20	0.351	(0.011)	74,739	0.127	(0.003)

Table 3: Yield of testing, i.e., probability of treatment among the tested (1), and cost per quality-adjusted life year (2), by ventile of model-predicted risk ($\hat{y}$ in a hold-out set of tested patients. Column 3 shows the probability of testing within each ventile of risk, in the entire hold-out set, i.e., tested and untested patients alike (bin thresholds are defined in the tested pool to maintain comparability, so "ventiles" do not have equal numbers of observations in the overall population).

Risk	Testing Rate			Yield		
Quartile	ECG Abnormal	ECG Normal	p	ECG Abnormal	ECG Normal	p
	(1)	(2)	(3)	(4)	(5)	(6)
1	.0351	.0192	< 0.001	.0746	.0294	0.038
	(.0073)	(.0057)		(.0333)	(.0285)	
2	.0562	.0268	< 0.001	.1259	.0672	0.049
	(.0090)	(.0061)		(.0405)	(.0420)	
3	.0873	.0482	< 0.001	.2107	.1048	0.006
	(.0107)	(.0084)		(0469)	(.0579)	
4	.1560	.1121	< 0.001	.3754	.2184	0.003
	(.0147)	(.0158)		(.0543)	(.0919)	
	ECG No STE	ECG STE	p	ECG No STE	ECG STE	p
1	.0280	.0000	< 0.001	.0580	.1250	0.610
	(.0050)	(.0000)		(.0253)	(.2450)	
2	.0401	.1290	< 0.001	.0848	.6667	< 0.001
	(.0054)	(.0841)		(.0277)	(.2469)	
3	.0669	.2542	< 0.001	.1679	.7273	0.003
	(.0069)	(.1127)		(0378)	(.2760)	
4	.1391	.4225	< 0.001	.3169	.8421	< 0.001
	(.0116)	(.1203)		(.0478)	(.2108)	

Table 4: Testing rate and yield of testing, by quartile of model-predicted risk $\hat{y}$ in electronic health records, as a function of two important findings on electrocardiograms (ECGs) done in the ED: whether the ECG was judged to be "normal" (reassuring), or whether the ECG shows signs of ST-elevation (highly concerning for heart attack). Columns 1-2 compare testing rates by ECG findings, and Columns 4-5 compare yield among the tested by ECG findings (with Columns 3 and 6 showing p values for the differences, respectively). The top panel shows data separated by whether the ECG was judged to be normal by the cardiologist or not; the bottom panel shows data separated by whether ST-elevation, which is concerning for heart attack, is present or not. All models are specified as linear probability models.

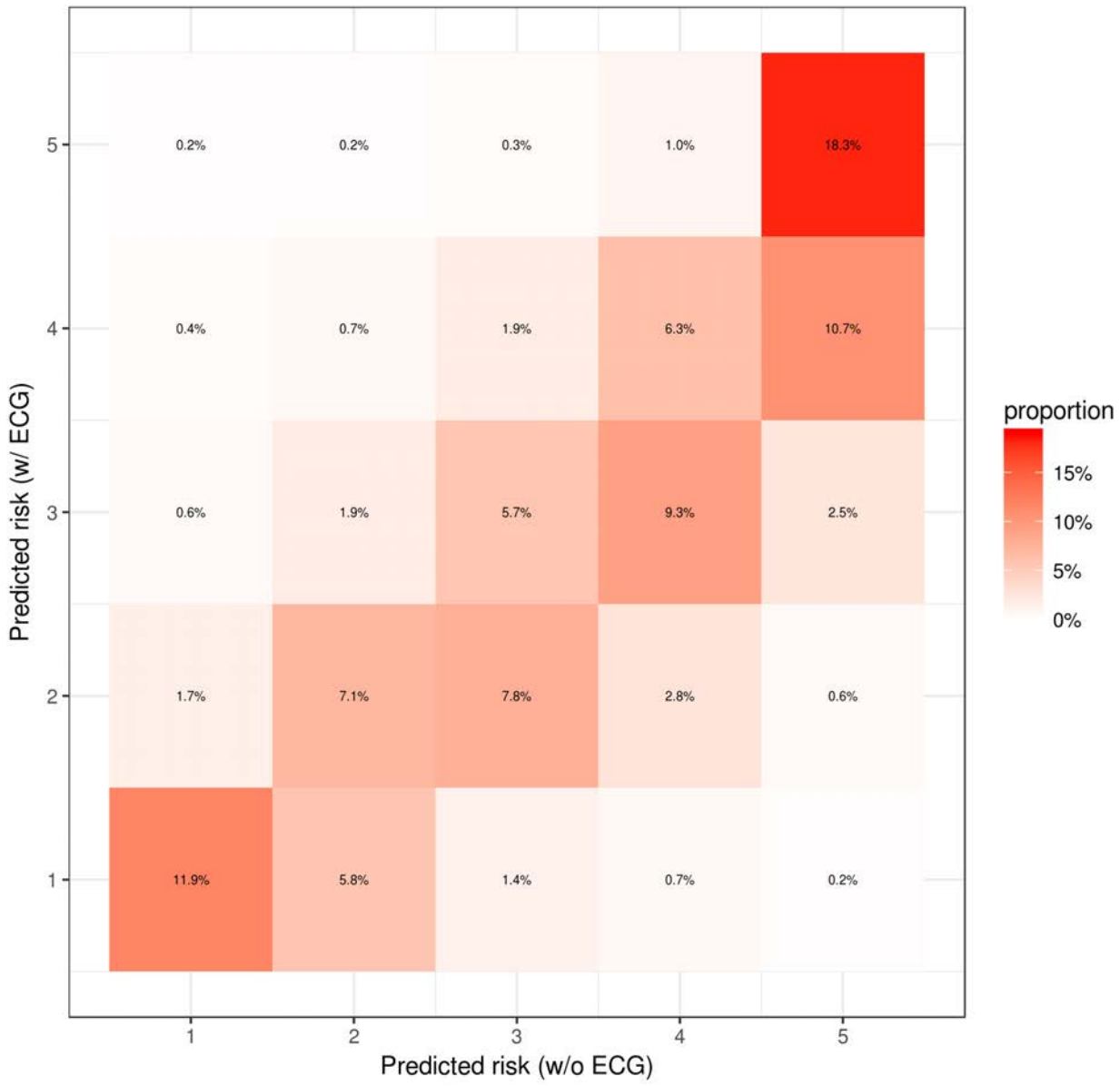

Figure 1: Comparison of model-based risk estimates, with (y-axis) and without (x-axis) incorporation of learned ECG waveform features. Bins are constructed in absolute $\hat{y}$ space on the original predictor.

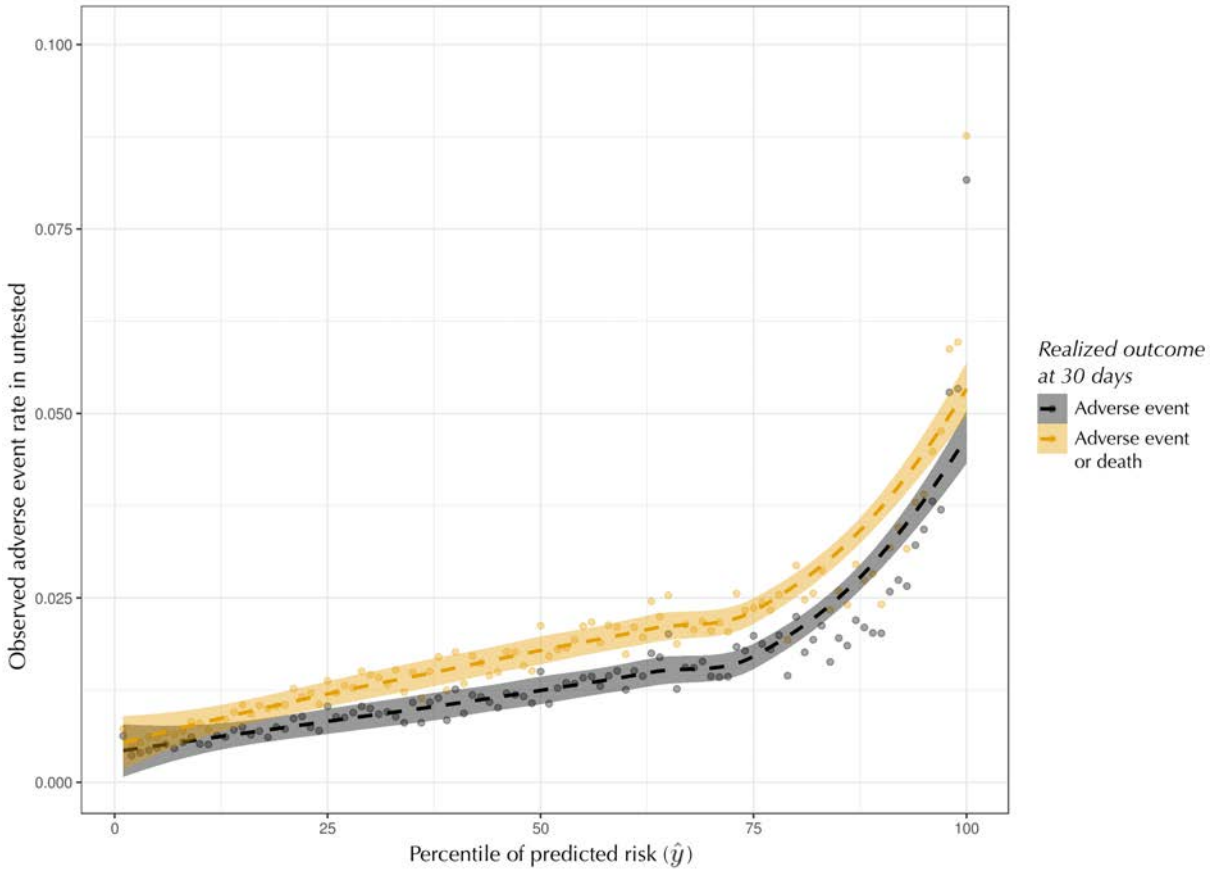

Figure 2: Rates of adverse events (return for heart attack or revascularization, cardiac arrest) and death, by percentile of model-predicted risk ($\hat{y}$) in a hold-out set of untested patients.

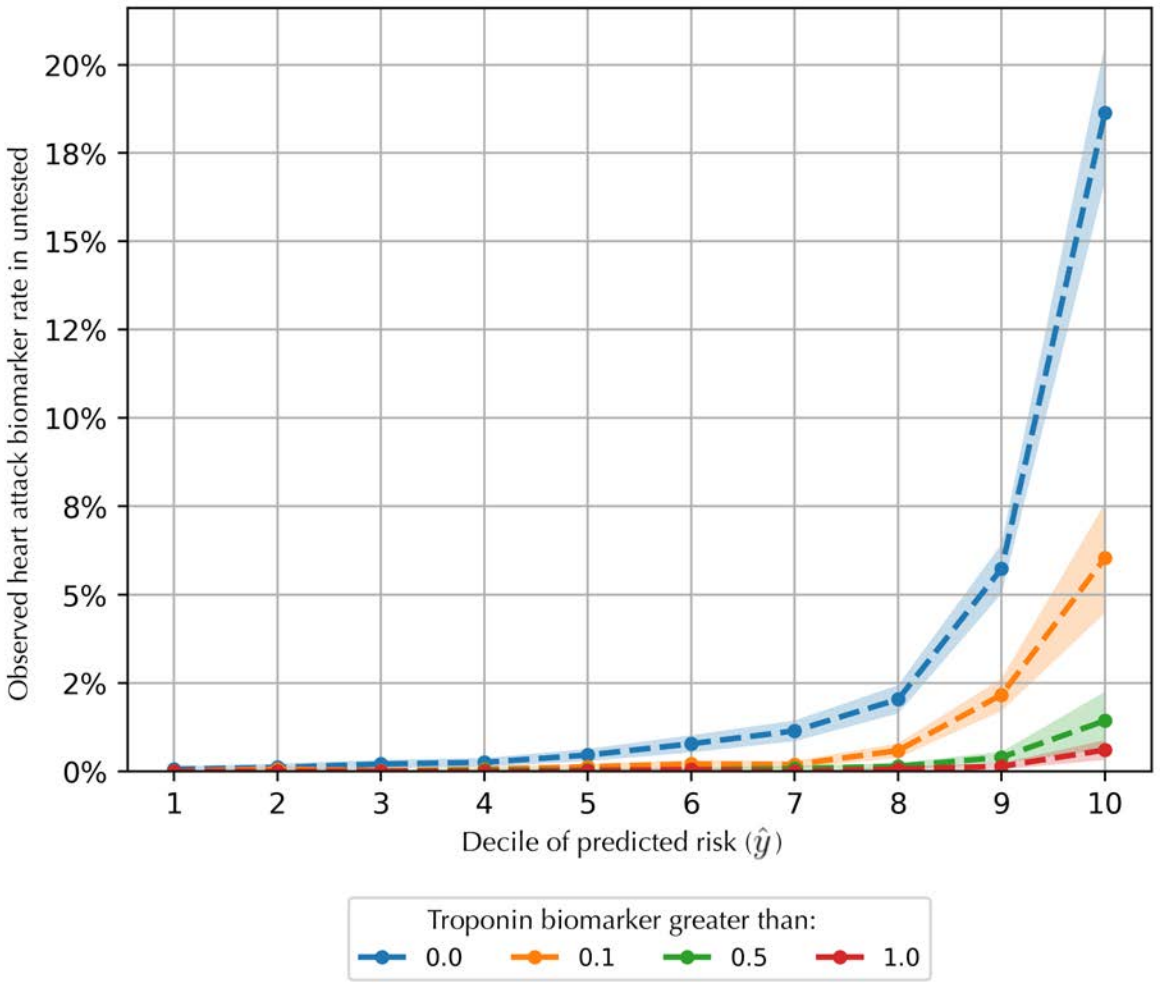

Figure 3: Rates of biomarker-confirmed heart attack, measured by maximum cardiac troponin level in a given patient in the 30 days after visits, by decile of model-predicted risk ($\hat{y}$) in a hold-out set of untested patients.

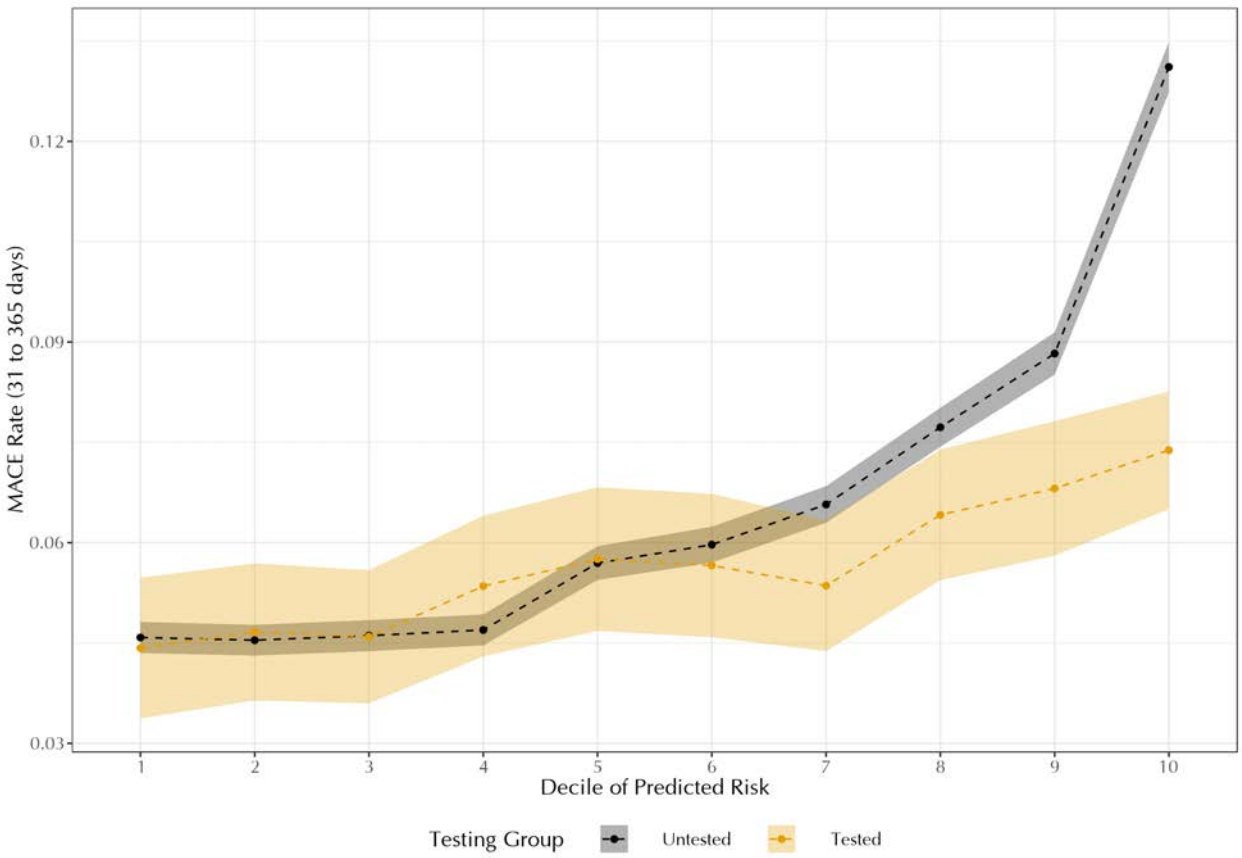

Figure 4: Rates of adverse events (return for heart attack or revascularization, cardiac arrest) in the year after visits (excluding the first 30 days), by decile of model-predicted risk ($\hat{y}$). The left panel shows rates for tested vs untested patients in the hold-out set, and the right panel shows the difference with 95% confidence intervals.

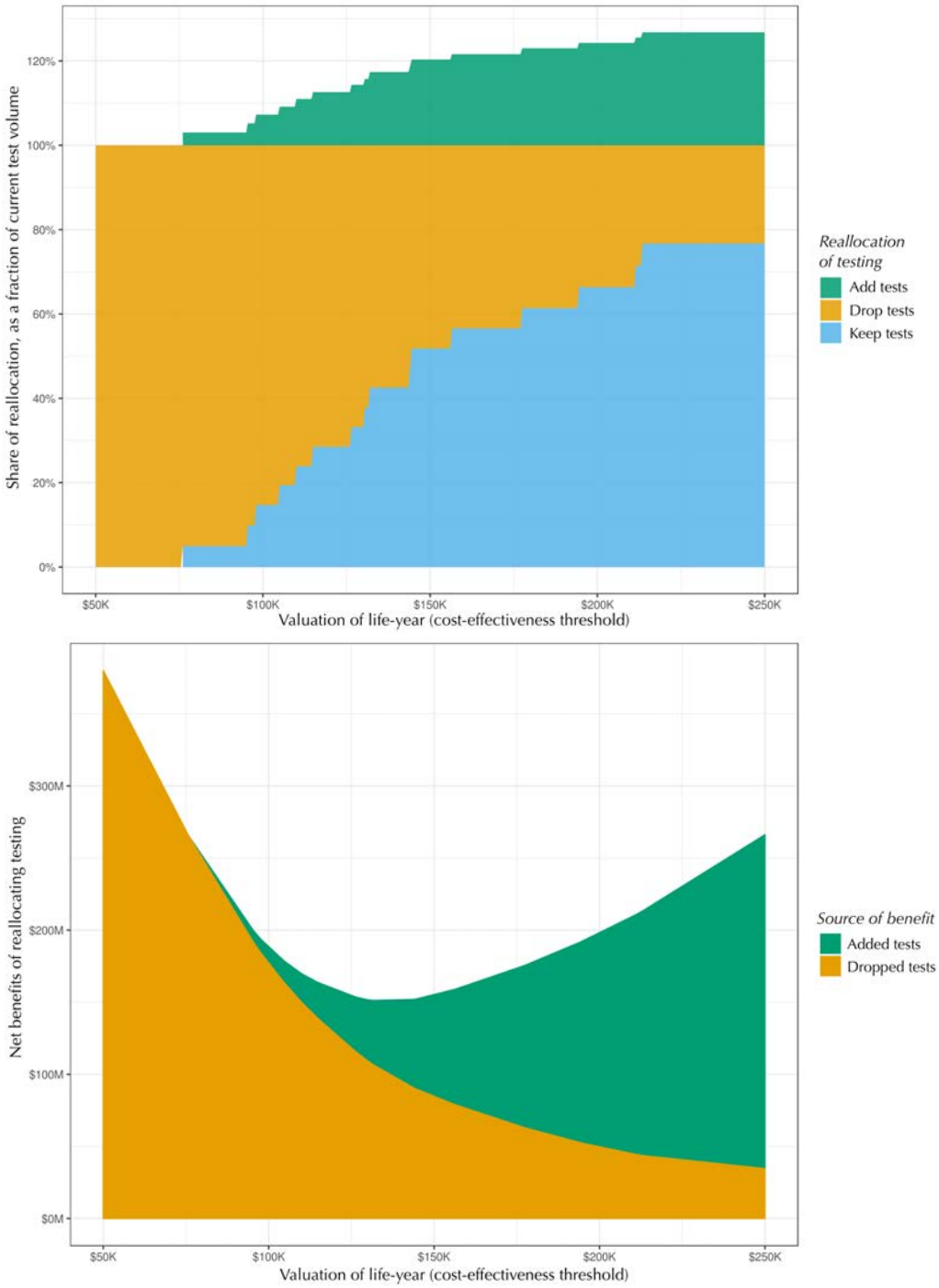

Figure 5: Net over- vs. under-testing at specific cost effectiveness thresholds. Top panel: Fraction of tests currently done that we would keep or drop; and how many of the untested we would choose to test (using a lower bound based on realized heart attack). Bottom panel: Net benefits of reallocating testing according to risk, combining surplus from life years saved, as well as the costs of testing. Note: these numbers show dollar values calculated in the holdout set, across 4.5 years of data, in the 20% random sample of claims.

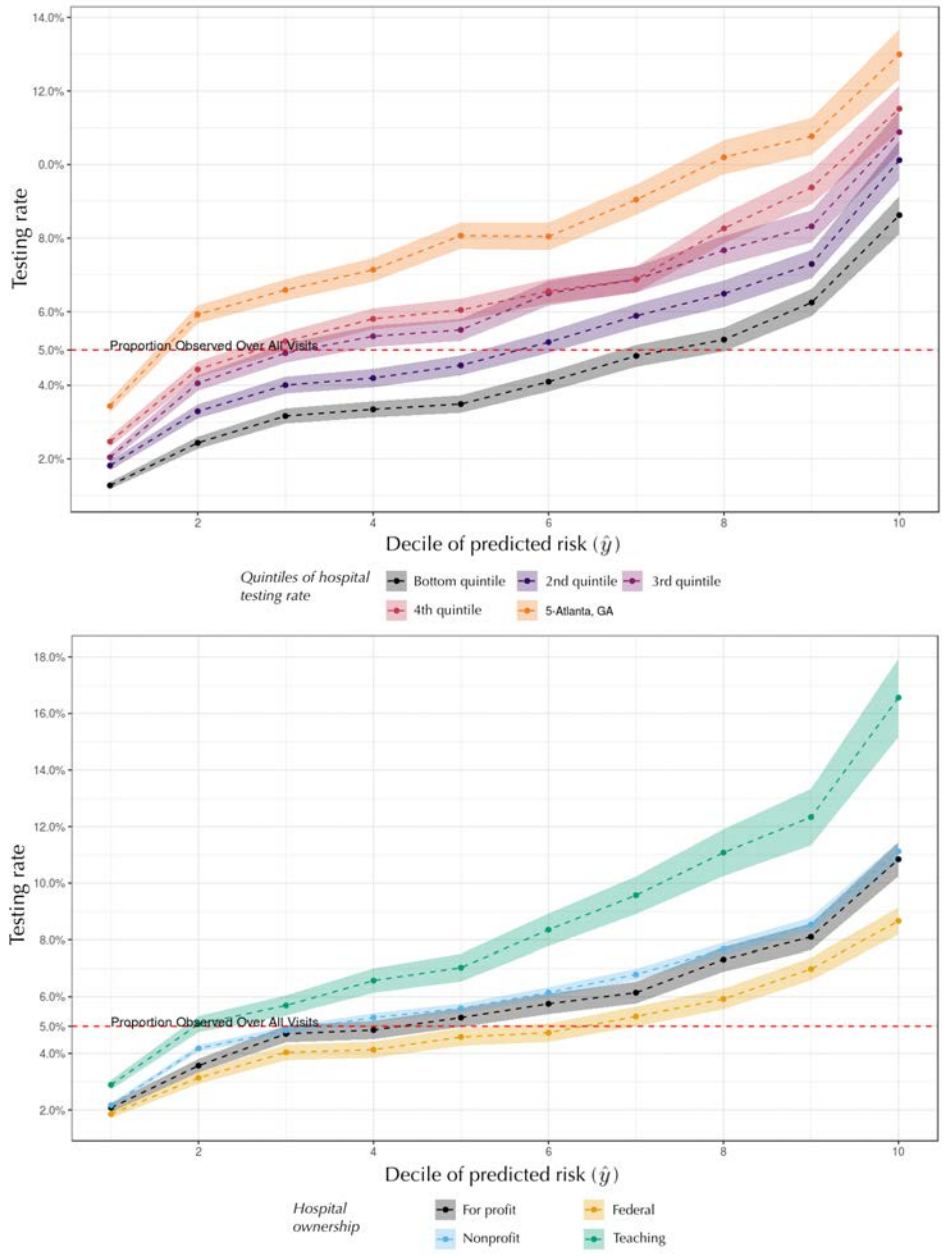

Figure 6: Difference in testing rates across hospitals, by distribution of predicted yield $\hat{y}_i$. Top panel: by quintiles of testing rate and labeled by largest metropolitan area. Bottom panel: by ownership structure. Hospitals with unknown ownership, and those missing Dartmouth data on HRR and spending are excluded.

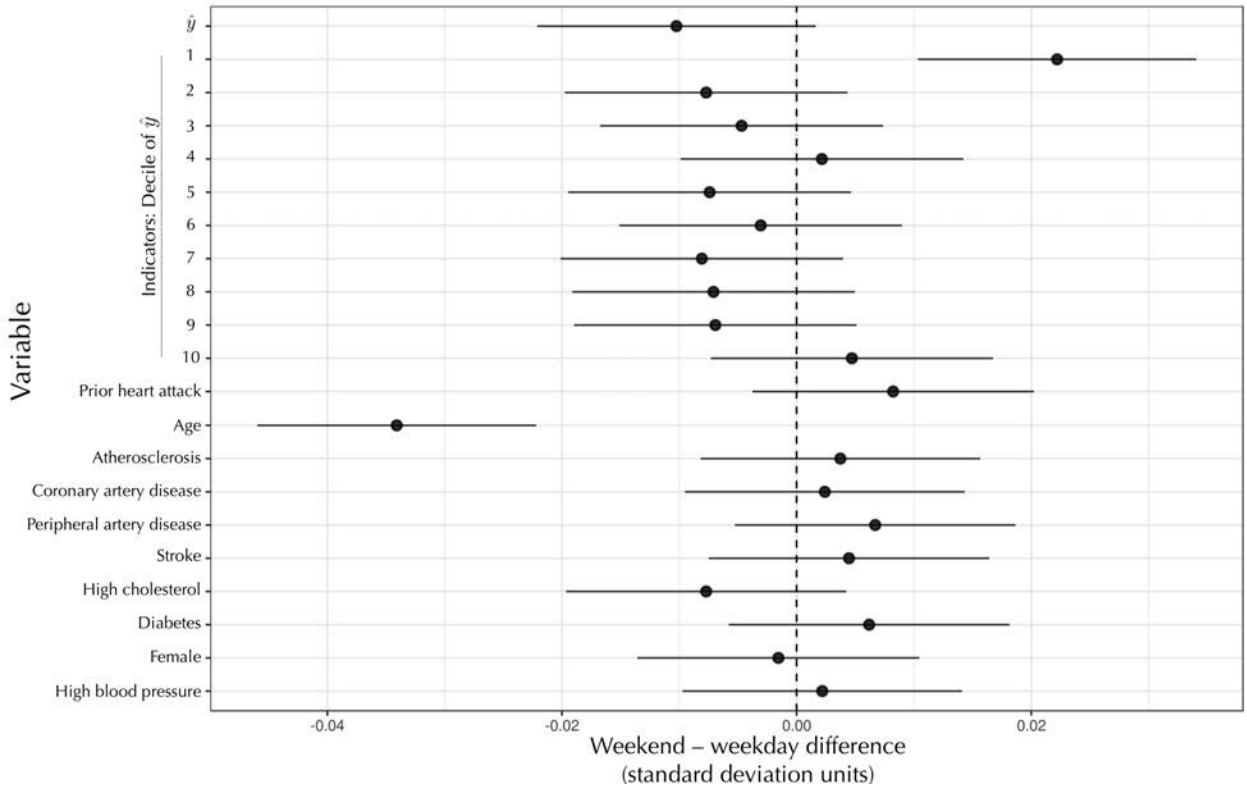

Figure 7: Characteristics of patients visiting on weekend vs. weekdays: Balance on $\hat{y}_i$ (first row), indicators for decile of $\hat{y}_i$ (rows 2-11), and other relevant observables. All differences are conditional on geography (i.e., hospital referral region) and year.

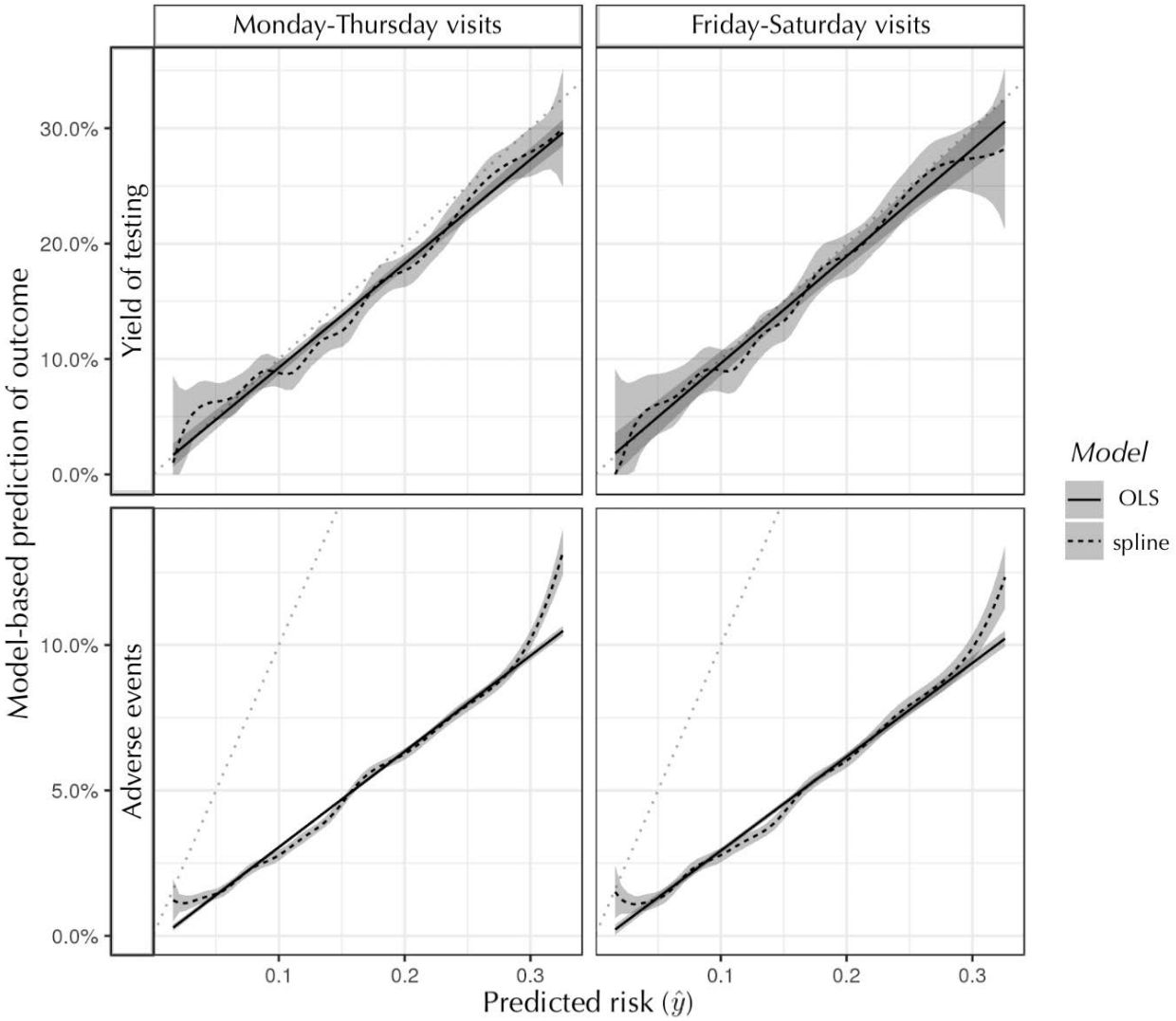

Figure 8: Realized yield of testing vs. decile of predicted yield $\hat{y}_i$ in the tested, weekend vs. weekday. Because the samples are small, we fit this using (1) OLS of the outcome (yield or adverse event) on an indicator for the weekend, $\hat{y}_i$, and an interaction term; and (2) a spline version of the same model, with 11 knots in $\hat{y}_i$ and the interaction term to verify calibration. We omit the top and bottom 2.5% of $\hat{y}_i$ in order to be able to consistently fit the spline at the tails.

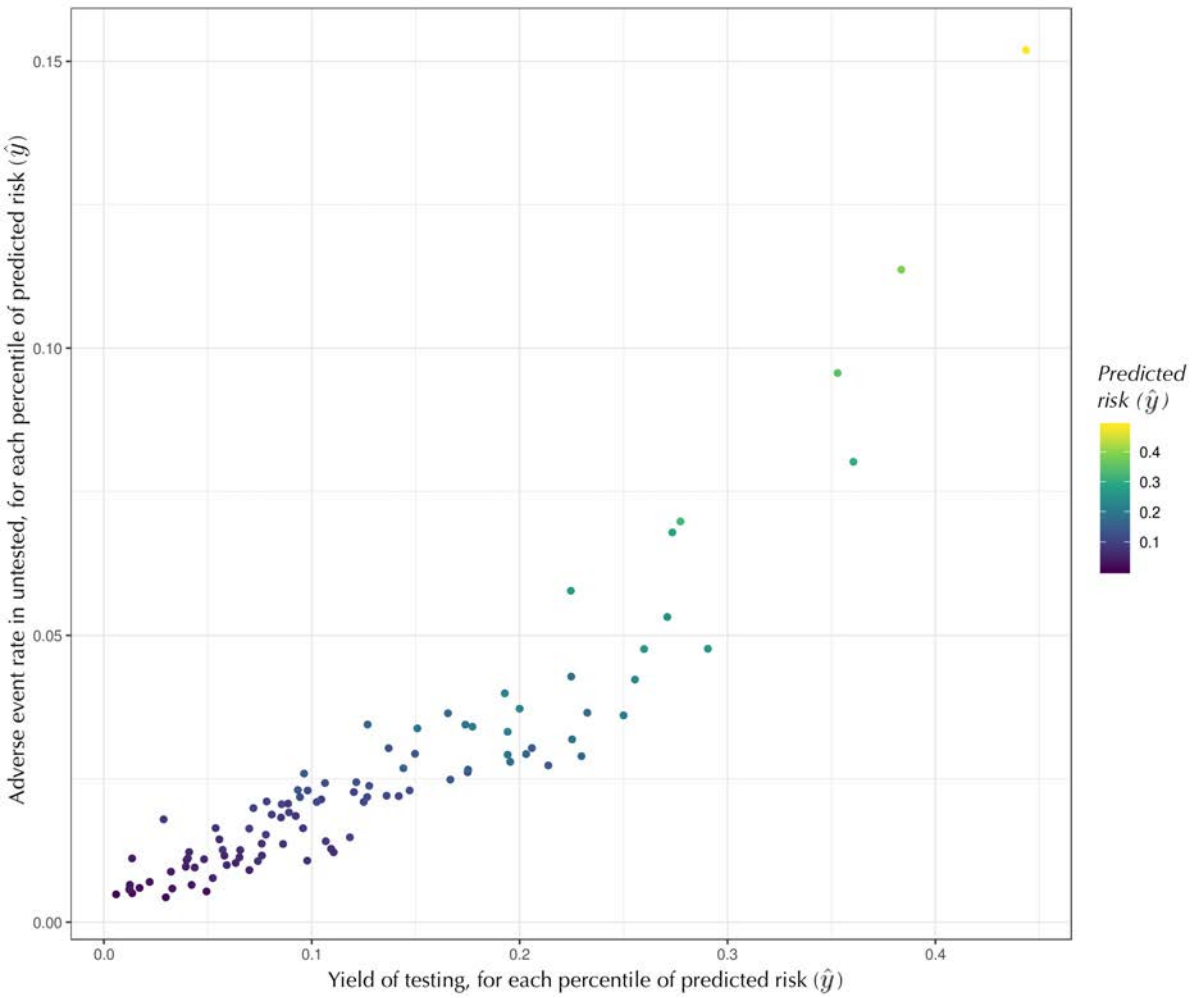

Figure 9: The figure first divides all-comers into percentile bins of $\hat{y}_i$ in the weekend vs weekday sample. It then shows, for each bin, the realized yield in the tested (x-axis) and the realized adverse event rate in the untested (y-axis), to demonstrate the relationship between the two outcomes.

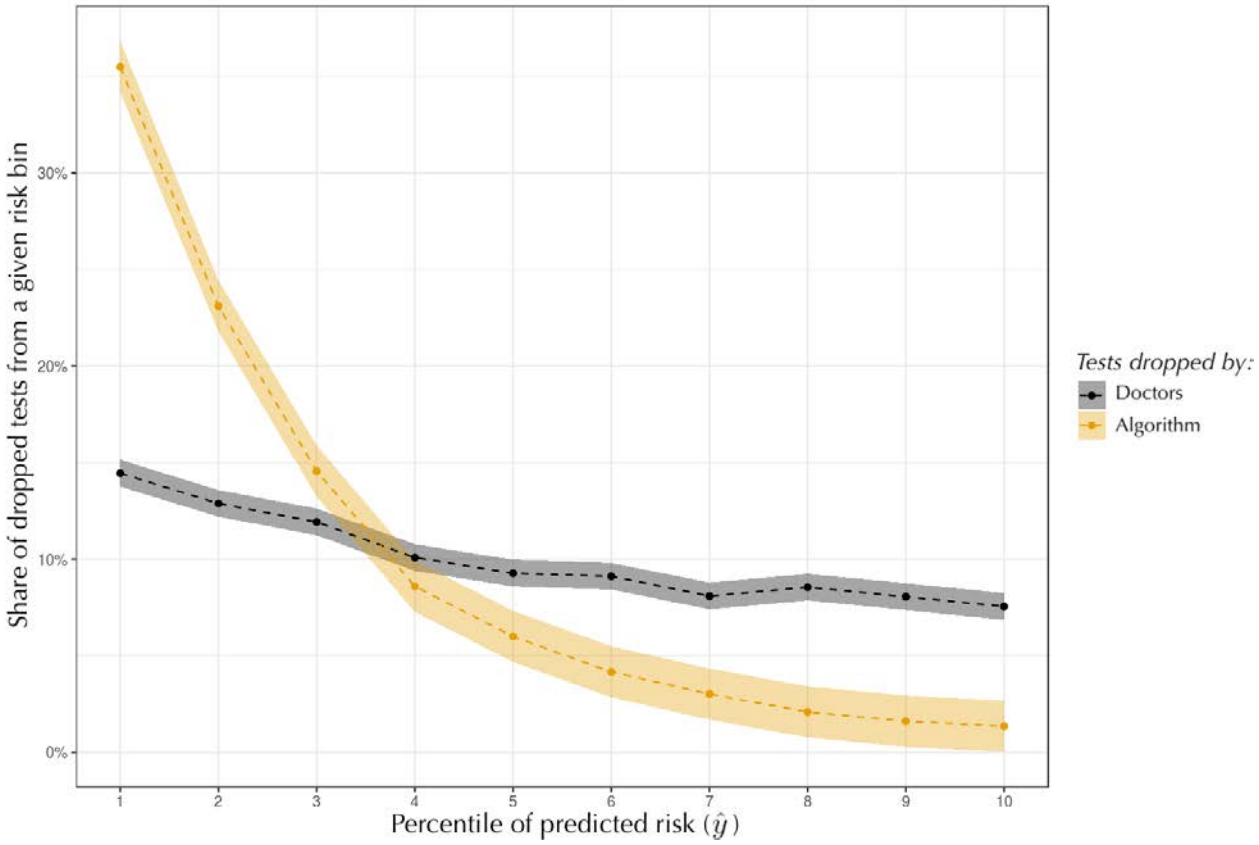

Figure 10: Risk distribution of marginal patients: Difference in testing rates, weekday vs weekend, by distribution of predicted yield $\hat{y}_i$, comparing observed doctor testing decisions vs simulated algorithm "decisions" (conditional on geography and year). Note: Decile bins are defined using cutoffs from tested patients, so these "deciles" in the overall population do not have equal size.

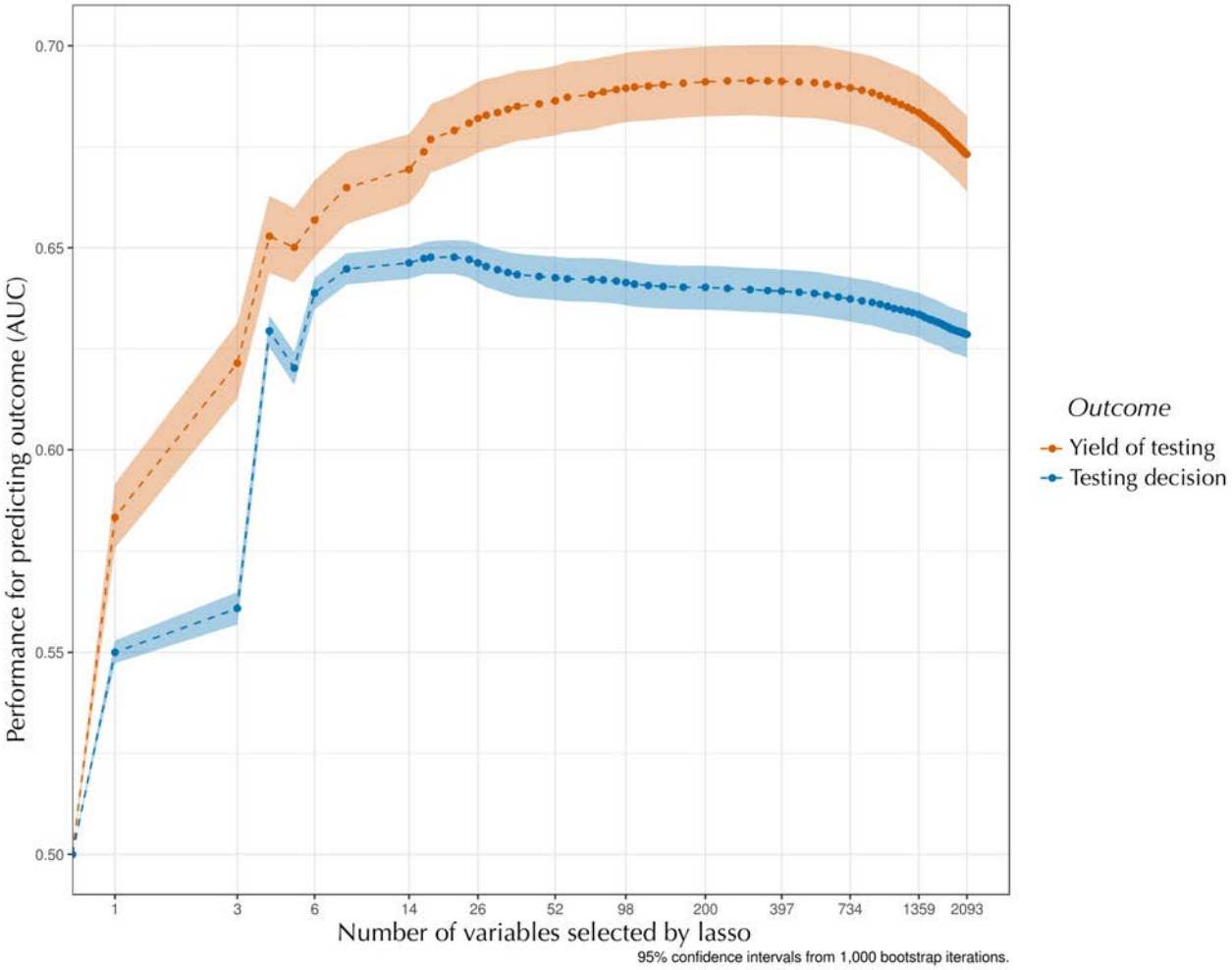

Figure 11: Predictive performance of simplified models of true risk. Models from the lasso path (x-axis: number of non-zero variables in the model) are used to predict true risk (top line) and doctors' testing decisions (bottom line). Performance is measured by area under the curve (AUC).

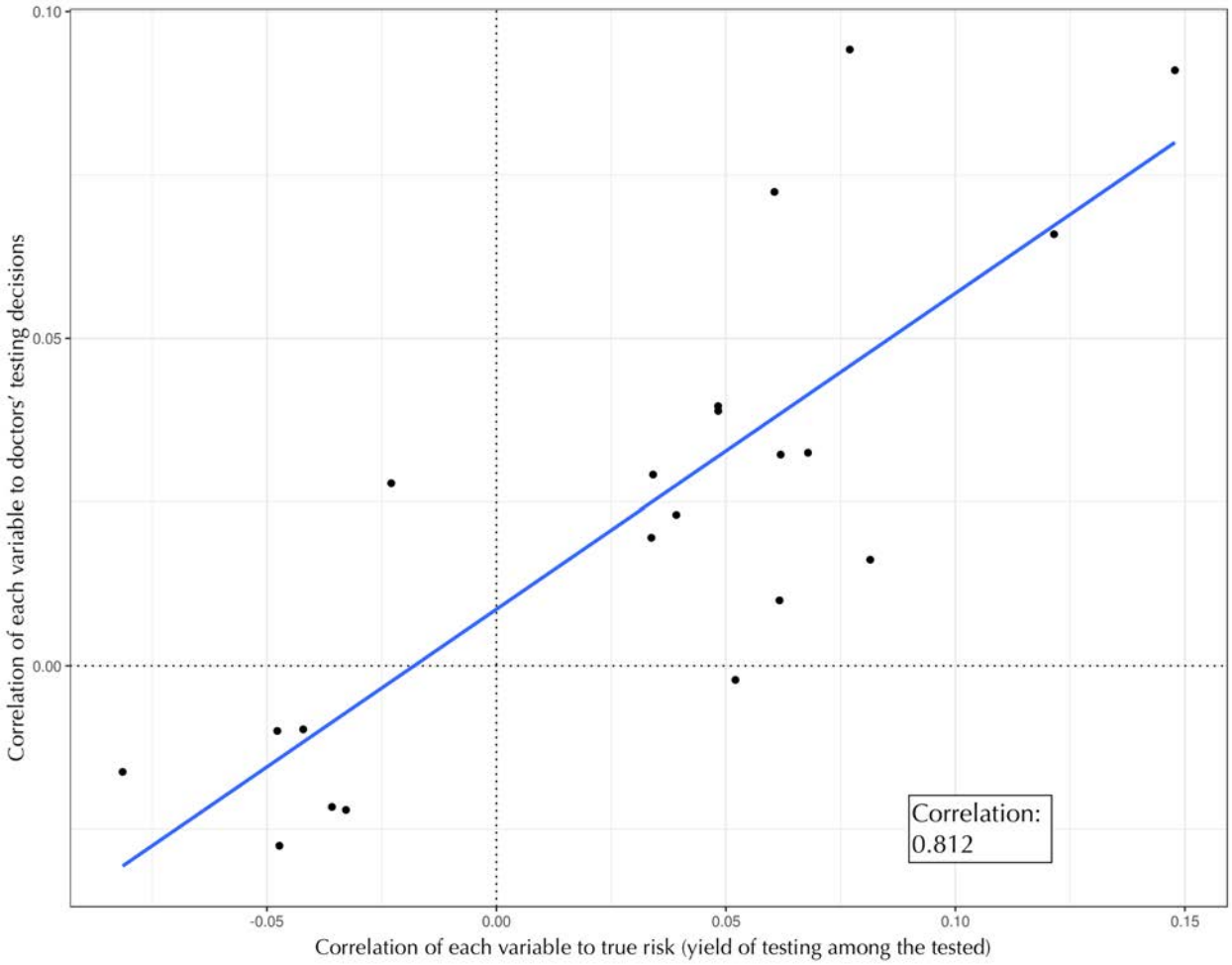

Figure 12: For the 21 variables included in the model that best predicts doctors' testing decisions, univariate correlations of each X with the testing decision (y-axis) and true risk (x-axis).

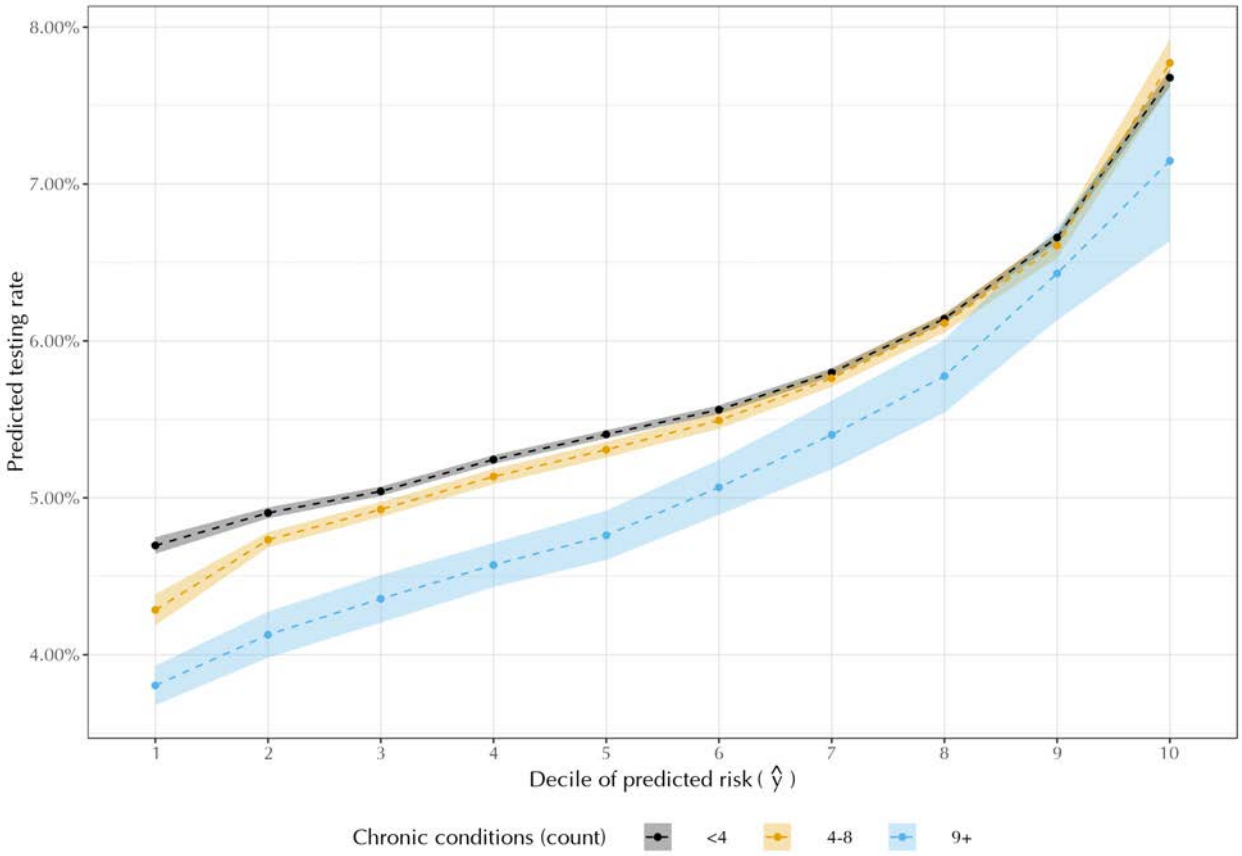

Figure 13: Rates of testing (predicted) for patients of different complexity, by decile of model-predicted risk ($\hat{y}$.) Note: the y-axis shows mean predicted testing rate, based on patient observables, rather than actual testing rate, since the latter was visually noisy in the smaller comorbidity-risk bins.

	Testing Rate			Yield		
	(1)	(2)	(3)	(4)	(5)	(6)
Predicted risk	0.064	0.067	0.056	1.005	1.00	1.005
	(0.012)	(0.012)	(0.012)	(0.067)	(0.066)	(0.067)
Chronic illnesses						
Count	-0.001			0.001		
	(0.0002)			(0.001)		
$4-8$		-0.004			0.002	
		(0.001)			(0.007)	
9+		-0.014			0.003	
		(0.002)			(0.017)	
Cardiovascular			0.006			0.002
			(0.001)			(0.003)
Other			-0.004			0.0004
			(0.0003)			(0.002)
Constant	0.046	0.043	0.041	-0.016	-0.012	-0.017
	(0.002)	(0.001)	(0.002)	(0.010)	(0.008)	(0.010)
N	166,314	166,314	166,314	7940	7940	7940

Table 5: Testing rate and yield as a function of risk and chronic illnesses. Columns 1-3 show how testing rate varies with the number of chronic illnesses a patient has. Column 1 uses a simple count of the number of illnesses, while Column 2 divides patients into three roughly equal-sized bins based on the number of illnesses. Column 3 divides illnesses into cardiovascular (hypertension, heart failure, etc.) and other. Columns 4-6 show the same analyses, but with the yield of testing (in the tested) on the left-hand side, to ensure that differences in testing rate are not a function of physician private information. All analyses are performed in patients without prior diagnoses of heart attack or stroke, in whom a new diagnosis is likely to matter the most (the Supplement shows similar results for the overall population). Notes: All models are specified as linear probability models. Sample mean test rate: 0.0195. Sample mean yield: 0.0895.