NBER WORKING PAPER SERIES

WHO IS TESTED FOR HEART ATTACK AND WHO SHOULD BE: PREDICTING PATIENT RISK AND PHYSICIAN ERROR

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

Authors contributed equally. We acknowledge financial support from 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, 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.

Who is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error

Sendhil Mullainathan and Ziad Obermeyer

NBER Working Paper No. 26168
August 2019
JEL No. C55,D8,D84,D9,I1,I13

ABSTRACT

In deciding whether to test for heart attack (acute coronary syndromes), physicians implicitly judge risk. To assess these decisions, we produce explicit risk predictions by applying machine learning to Medicare claims data. Comparing these on a patient-by-patient basis to physician decisions reveals more about low-value care than the usual approach of measuring average testing results. It more precisely quantifies over-use: while the average test is marginally cost-effective, tests at the bottom of the risk distribution are highly cost-ineffective. But it also reveals under-use: many patients at the top of the risk distribution go untested; and they go on to have frequent adverse cardiac events, including death, in the next 30 days. At standard clinical thresholds, these event rates suggest they should have been tested. In aggregate, 42.8% of the potential welfare gains of improving testing would come from addressing under-use. Existing policies though are too blunt: when testing is reduced, for example, both low-value and high-value tests fall. Finally, to understand physician error we build a separate algorithm of the physician and find evidence of bounded rationality as well as biases such as representativeness. We suggest models of physician moral hazard should be expanded to include 'behavioral hazard'.

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

Physician judgment plays a key role in all aspects of health care: in the course of roughly 3 million patient encounters every day in the United States, physicians make decisions that affect both the quality and cost of care. Here we study a common and representative physician judgment: the decision to test for heart attack. Consider a patient complaining of chest pain or shortness of breath. While such symptoms are consistent with a heart attack, they might also reflect more benign conditions, such as acid reflux. Timely diagnosis of heart attack would be very valuable: proven treatments can extend life and prevent serious complications. A highly diagnostic test exists, but it is invasive and expensive, and so must be applied selectively. We empirically study these testing decisions. How are physician judgments formed? Are the right people tested? Can decisions be improved upon? These questions are useful to answer in part because of the importance of diagnosis for heart attack, which remains one of the leading causes of death in the US (Murphy et al. 2018) and globally (Roth et al. 2018). But also because the approach we take to answer them could apply more broadly: not only to other testing decisions – such as laboratory tests, imaging, cancer screening – which account for a large fraction of health care, but also to analogous decisions in other sectors.

We begin by noting that the decision to test is implicitly a prediction problem. Since the value of a test is to provide new information, the decision to test depends on its expected yield – the likelihood of a positive test.[1] Physicians must aggregate available information – symptoms, physical examination, medical history, prior testing, etc. – to form an individual-level estimated probability of a positive test. So even though the physician ultimately allocates treatments, grounded in a causal model of disease, her *predictive* model also plays a central role, via the testing decision (Kleinberg et al. 2015). Since a physician's predictions are formed on a patient-by-patient basis, simply looking at her average performance – overall testing rate or yield – would be a weak way to measure the quality of her predictive model. Instead we require tailored, individual-level risk predictions against which to compare individual testing decisions. High dimensional prediction tools (machine learning) are ideally suited to form such estimates. This is especially true in the medical context, which as many have noted, provides unusually rich input data on patient characteristics.[2] We build a risk predictor using nationwide Medicare claims data to both assess the overall quality of decision making, and to gain specific understanding into the nature of physician errors.

In health economics, the dominant model of physician decision making emphasizes incentives and moral hazard: since insurers typically pay per test, doctors have an incentive to over-test. In this view, doctors form a risk prediction, and test all patients above a certain threshold; but because of poorly aligned incentives they set the threshold too low, leading to wasted tests for patients with predictably low risk. The usual way the literature has documented such inefficiencies is to examine average yield of testing, which is then translated into a cost effectiveness estimate that allows for normative evaluation of whether testing is a worthwhile activity. Applying the usual

[1] We make several specific assumptions here, which are discussed in detail later on. By a 'positive test' we mean one that changes the treatment course. We thus equate information value of the test to its effect on decisions. So we abstract entirely from any psychological benefits of knowing that one does not have a heart attack. We also assume that downstream treatments are effective and useful, which in this context is a reasonable assumption as we discuss below. We will also assume that prior uncertainty is such that the highest ex ante probability of a positive test is still sufficiently far from 1 that testing still adds information.

[2] For machine learning applications in medicine, Rajkomar, Dean, and Kohane 2019 provide a helpful overview. Recent notable examples include Ghassemi et al. 2014 and Rajkomar et al. 2018 who use these tools to predict outcomes like mortality or readmission, and Miotto et al. 2016 who predict a variety of future diagnoses.

cost-benefit assumptions from the literature to our data, we find the average test has an implied cost effectiveness of $135,859 per life year.[3] From this we might conclude that testing as a whole is barely cost effective, or slightly ineffective (depending on which threshold we use, typically $100-150,000 in the US (Neumann, Cohen, and Weinstein 2014). Yet, if the problem is that doctors set too low a threshold, we should focus on the marginal test near the threshold, not the average test. Individualized algorithmic risk predictions, by characterizing the entire risk distribution, allow us to more precisely measure the extent of over-testing, and find more striking inefficiencies than the average suggests. For example, in the 10% of tested patients with the lowest predictable risk, testing costs $616,496 per life year; and at a threshold of $150,000 per life year, 52.6% of tests would be cut.

A closer examination of the full risk distribution reveals another curious fact, at the other end of the risk spectrum. Unsurprisingly, testing these patients is highly cost effective: in the 10% of tested patients with the highest predictable risk, testing costs $82,621 per life year, well within even stringent bounds for cost effectiveness. Yet despite this apparent value, a surprisingly large fraction of these patients go untested. This raises the possibility that physicians do not just over-test but also *under*-test, as suggested in earlier empirical work (Chandra and Staiger 2007; Abaluck et al. 2016).[4] Of course, we must be careful in drawing such conclusions: we do not know what tests might have yielded for these untested patients. Any effort to document under-testing must grapple with basic selection bias, particularly when driven by unobservable variables that influence the testing decision but do not appear in existing datasets. For example, physicians question and examine the patient, obtain a variety of tests, and generally elicit a great deal of information that is at best imperfectly measured, or simply not recorded at all. As a result, when physicians fail to test a patient whom the algorithm flags as high risk, it could be an error – or it could be an effective use of available data to which the algorithm is not privy. We illustrate the scale of this problem using a unique dataset of electrocardiogram (ECG) tracings that are not available in claims data, and rarely available even in electronic health records, to show how such unobservables can overstate the extent of human error in machine learning applications.[5]

To address this bias, we exploit the panel nature of our data: we can follow untested patients and measure their eventual outcomes. If high risk untested patients actually had undiagnosed heart attacks, consequences should manifest over time.[6] Precisely because these patients were untested

[3]Our cost effectiveness estimates incorporate all direct costs of testing in the Medicare population, including hospital or observation stays and physician fees, and benefits derived from applying treatment effects from the literature (Amsterdam et al. 2014; Bavry et al. 2006) to standard disability weights and life expectancy for heart disease (Mahoney et al. 2002). We include costs and life years lost due to disability from complications of catheterization. We do not include downstream costs (e.g., biopsy of a lung nodule discovered on perfusion imaging) or benefits (e.g., early diagnosis of otherwise fatal cancers) that may arise from testing.

[4]For example, Chandra and Staiger 2007 find what appears to be over- or under-use of interventions stemming from poor choices on the part of hospitals, and the structural model of Abaluck et al. 2016 shows counterfactual outcome distributions compatible with both over- and under-testing.

[5]This elementary selection bias pervades many machine learning applications to prediction policy problems. Following this literature's convention of naming dependent variables as 'labels' it is sometimes referred to as the selective labels problem (see e.g. Kleinberg et al. 2017). In medicine, a related problem arises when measured outcomes are selectively changed by treatment, as opposed to testing which selectively reveals them. For example, when physicians administer antibiotics and fluids to patients at risk of sepsis, these treatments make it hard to measure the extent of over-treatment (see Paxton, Niculescu-Mizil, and Saria 2013 for a thorough discussion).

[6]There is a long history of clinical research into the natural history of heart attack extending well into the modern era of medicine, in the absence of effective treatments until the early 1980s. These studies (e.g., trials comparing hospital vs. home-based management of heart attack and finding no benefit: Mather et al. 1976; Hill, Hampton, and

(and as we show, undiagnosed) their outcomes are not masked by treatment. We find that, in the 30 days after visits, untested patients in the highest-risk decile have strikingly high rates of major adverse cardiac events: 3.8% return to care, only to be diagnosed with heart attack or the cardiac arrest that results from it, or to have an urgent treatment intervention; an additional 1.5% drop dead. Looking back to their initial encounters, most of those with realized adverse events were sent home from the emergency department (55.1%) instead of hospitalized, implying that doctors under-estimated their risk; and a majority (60.4%) were given diagnoses (e.g., acid reflux) suspicious for missed heart attack.[7] To calibrate whether these rates imply under-testing, we turn to the clinical literature, which gives us thresholds for levels of risk that would mandate testing.[8] These thresholds are consistently far lower than the rates we observe, typically between 1 and 2%, as are surveys of doctors regarding an acceptable miss rate: 1% (Than et al. 2013).

Synthesizing these empirical results, we arrive at our first core finding: physicians simultaneously both over- and under-test. To calculate the relative magnitudes of each, we compare the current testing regime to one implied by the algorithm's risk scores. Doing so requires calculating a lower bound for the extent of under-testing, which we infer indirectly from adverse outcomes in the untested. We observe that while some predictably under-tested patients might not go on to experience adverse events, the minimum number of under-tested patients is the number who experienced adverse outcomes. We thus use the adverse event rate to form a conservative lower bound, and find that, at a value of \$150,000 per life year we would add in 17.9% of tests for high risk patients not currently tested, while dropping 52.6% of tests doctors currently choose to do – on net, reducing testing by 34.7%. But because the net benefits of reducing over- and under-testing are different, we find that 42.8% of the net benefits of reallocating testing in this way would come from increasing testing of the high-risk untested, generating \$228.0 million in annual surplus from life years saved; 57.2% (or \$304.7 million) come from eliminating wasteful tests.

A model where physicians simply use too low a risk threshold cannot account for simultaneous over- and under-testing: physicians must also *mis-predict* in a way that produces faulty rank ordering of risk. Overlooking mis-prediction can lead to faulty interpretations and policies. For example, research has documented wide variations in testing rates across hospitals, but little variation in outcomes: hospitals that test much more seem to have no commensurate health benefits. The common interpretation is that these additional tests are all low value, and the common solution is to encourage high testing hospitals to test less. Considering the full distribution of risk, by contrast, paints a more complex picture: high-testing hospitals do test more low risk patients – but they also test more high risk patients. We find a similar pattern using variation linked to hospital ownership. This suggests that encouraging physicians to cut tests would reduce over-testing, but also exacerbate under-testing. We solidify this finding by using quasi-random variation to more precisely quantify the margin: since patients are typically observed overnight before testing, but hospital staffing is decreased on weekends, it is more difficult to test patients who come in on *the day before* a weekend day. Empirically, we find patients who come in on Fridays and Saturdays are 19.8% less likely to be tested (3.95% vs. 4.93%, $p < 0.001$, despite looking similar on observables). This lets us confirm both the predictive accuracy of our model in a setting with low unobserved

Mitchell 1978) describe in detail the fate of untreated patients: recurrence of heart attack, arrhythmias, and death.

[7] These are codified in a substantial literature on diagnostic error, most recently Wilson et al. 2014.

[8] 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).

differences, and verify that marginal patients in this context too are drawn from both predictably high and low risk groups. We view these results as evidence that policy models of physician decision making should incorporate error, analogous to 'behavioral hazard' models of patient decision making (Baicker, Mullainathan, and Schwartzstein 2015; Brot-Goldberg et al. 2015; Handel and Kolstad 2015).

Our second core finding is an answer to the natural question: what then drives physician mis-predictions? To see if machine learning tools can help us understand human error, we build another algorithm, this time to predict physician choices directly. By contrasting the algorithmic model of the physician with the model of true risk, we can partly characterize errors.[9] To put our findings in context, we differentiate between two categories of behavioral models. The first, going back at least to Herbert Simon (Simon 1955), emphasizes bounded rationality, focusing on the finiteness of cognitive resources such as attention, memory or computation (Gabaix 2014; Sims 2003; Gabaix 2017; Mullainathan 2002a; Bordalo, Gennaioli, and Shleifer 2017). In our setting, for example, physicians may not attend to, process, or mentally represent the entire set of patient characteristics they face, and instead rely on a simpler model of risk. The second view emphasizes that, even in the set of variables used for decision making, people make systematic mistakes. For example, probabilistic errors – such as representativeness and anchoring – arise when human judgment deviates from the rules of probability (Tversky and Kahneman 1974).

Empirically, we do find evidence for boundedness: a simpler (more regularized) model of true risk predicts physician judgments better than more complex (less regularized) models of risk, suggesting physicians rely on too simple a model. Interestingly, the variables used in the simple risk model appear to be weighted approximately correctly: a given variable's weight for predicting testing and its weight for predicting risk are highly correlated (0.812). At the same time, this is not a perfect correlation, and by examining variables that predict testing above and beyond true (algorithmic) risk, we discover two suggestive instances of systematic error. First, we find evidence that physicians over-react to demographic risk. They over-test patients from high risk demographic groups and under-test those from low risk groups. We interpret these results as consistent with representativeness broadly (Kahneman and Tversky 1972) and a recent model of stereotyping specifically (Bordalo et al. 2016). Second, we find that patients with a history of pneumonia are less likely to be tested for heart attack conditional on risk, and therefore more likely to suffer adverse events. Since some of the symptoms of heart attack can be 'explained' by pneumonia, the most parsimonious explanation here (pneumonia) is also a misleading one, an error we call 'Occam's eraser,' which relates to findings from a large clinical literature on diagnostic error.[10]

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 in the coronary arteries supplying it. This leads to damage or death of a patch of heart muscle, which has both immediate consequences, e.g.,

[9] This approach builds on a long history of research comparing clinical vs. actuarial decision making to gain insight into physician decision making (e.g., Ægisdóttir et al. 2006; Dawes, Faust, and Meehl 1989; Elstein 1999; Redelmeier et al. 2001). Our approach differs slightly by explicitly building an algorithmic model of physician judgment.

[10] For helpful overviews, see IOM 2015; Croskerry 2002, with specific examples in Graber, Franklin, and Gordon 2005; Schiff et al. 2009; Singh et al. 2013.

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.[11]

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).

Since simple tests done in the emergency setting (e.g., electrocardiograms, troponin testing) are often unrevealing or ambiguous, 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, 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 squirts 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. These are both direct – thousands of dollars for stress tests and tens of thousands for catheterization – and indirect, arising from the need for overnight observation and monitoring before testing can proceed. There are also 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 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).

Studies of testing across a variety of different medical settings have documented the same key empirical fact: low average yield. Nowhere is this more true 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

[11] 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.

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 this, 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. 2017) – 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, in which 62% of the items labeled as low-value were diagnostic tests (Morden et al. 2014).

2.2 Prediction Problem

To study this problem, we begin by formalizing the way in which the doctor makes her decision. In the case of testing, as with any other decision, we would want her to test patients above some threshold, at which 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 testing's 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, 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? So 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. This makes testing one instance of a more general 'prediction policy problem' (Kleinberg et al. 2015), requiring decision makers to synthesize complex inputs into predictions on an outcome.

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. Since tests produce health via the interventions they trigger for patients who are having heart attack,[12] 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 negative 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 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.[13]

Without these patients, it is reasonable to assume $Cov(T_i^+, \tau_i | X_i) \approx 0$ so that 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.

Using this model, we can now articulate the key predictions of moral hazard in this setting, and consider alternative models. Under moral hazard, when doctors form their predictions on the benefit of testing – i.e., the likelihood that a patient will test positive – they do not use the socially optimal testing threshold. Rather, they set the threshold too low, because they are incentivized to do so. Yes, they test those patients who are likely to benefit – but they also dip further and further down the risk distribution, testing lower and lower-risk marginal patients who are unlikely to benefit. When doctors are paid more to test more, and the vast majority of diagnostic tests come back negative, it is easy to see incentives at work. In addition to this over-testing, here is

[12] 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.

[13] 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.

another equally critical, but less well appreciated implication of this model: there should be little to no under-testing. Doctors set the threshold too low – but still test nearly all of the highest-risk patients who deserve to be tested, along with the marginal medium- and low-risk patients who are tested because of incentives.

Taking a more behavioral view of this decision making process, by contrast, might raise a number of questions about the particular risk prediction regime used by doctors. How do doctors predict risk? And are these judgments accurate? One could easily imagine a world where doctors, faced with the complex task of predicting which patients will benefit from a test, make mistakes. Even when incentives are aligned, physicians like other people may not conform to normative models of decision-making and may make judgmental errors (Thaler 1987). Medical decisions are particularly difficult and physicians' bandwidth might be taxed by high volumes and limited time (Balogh et al. 2016; Obermeyer and Lee 2017). As a result, they may use parsimonious ('sparse') models with lower accuracy (Simon 1955; Gabaix 2014). As the decisions involve significant uncertainty, they may also fall prey to any of the numerous well documented judgmental errors and biases in probabilistic judgment (Tversky and Kahneman 1974). At the extreme, if doctors' risk predictions were more or less random, average yield would be low (i.e., the base rate). So clearly mis-prediction, just like moral hazard, can also create low-value care. But critically, unlike moral hazard, mis-prediction would also imply the existence of large numbers of high-risk patients who are untested by doctors – not because of incentives, but because of errors in judgment.

Historically, glossing over the particulars of doctors' risk prediction regime was partly a technical limitation: it would have been difficult to say much in the absence of accurate individual-level estimates of risk. But this is an increasingly tractable problem today, thanks to novel tools for risk prediction, and the availability of rich, high-dimensional electronic data. Unlike the traditional methods of economics or epidemiology, predictive algorithms from the field of machine learning are designed for out-of-sample accuracy, and capable of handling the types of data found in electronic datasets. Of course, in making predictions and selecting patients for testing, doctors rely on factors that are both observable and unobservable. But at the end of the day, the yield of the test is observable, in the form of specific interventions that are either delivered to patients or not, meaning that the problem is tractable using data documented in the medical record or even insurance claims.

2.3 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). Even the results of stress testing and catheterization are complex to interpret: the tests provide a snapshot of the heart's electrical activity in response to increased demand, or a picture of the dynamics of blood flow in the coronary arteries – hardly binary variables.

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.

1. 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 conservatively choose treatment effects at the bottom of the range of those reported in the literature (reviewed in Amsterdam et al. 2014; Bavry et al. 2006). We note that, in the presence of over-treatment, the estimates of over-testing we present would represent a lower bound, since we would be giving the benefit of the doubt (and assigning a non-zero treatment effect) to all patients treated by doctors under the current treatment regime.

2. 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 (Schulman et al. 1999) – these biases have not been shown or suggested *after* testing, in the decision to treat patients conditional on test results. Indeed, this is one of the major advantages of anchoring the prediction on results of testing in tested patients.

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 (the 'selective labels' problem, analogous to potential outcomes in causal inference, which we discuss in more detail below). 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* ($T_i = 0$), 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? As a preliminary solution to this problem, we take 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 thus replicate this outcome (Y_i^U) in untested patients.

2.4 Data

National Medicare claims Using a nationally-representative 20% sample of Medicare claims data, we identified 20,059,154 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 those frail elders sent from a nursing home to the ED; 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 and treatment for heart attack 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 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 for use with Medicare claims as the Healthcare Common Procedure Coding System) is modified every year, with 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.[14]

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

[14]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.

all visits to a large, urban emergency department (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. 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 has stress tests and 1,668 had cardiac catheterization. Of these, 738 had revascularization procedures after the initial test.

2.5 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.[15]

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 visit to a cardiologist for high blood pressure); EHR data include all encounters recorded by 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),[16] 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, which can occasionally be back-dated. This results in two variables, describing occurrences over a recent and baseline time period, for each semantically-grouped diagnosis or procedure group. 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:

1. Revascularization intervention in the tested, $Y_i^T = 1|T_i = 1$

[15] We are necessarily brief in our description of machine learning methods; see Mullainathan and Spiess 2016 for a more thorough overview with references.

[16] As an example, this allows us to group low-level diagnosis and procedure codes (e.g., E018.2: Injury from activities involving string instrument playing) into broader clinically meaningful categories (e.g., E000-E999: External Causes Of Injury).

2. Adverse cardiac event in the untested, $Y_i^U = 1|T_i = 0$ (we discuss this outcome in more detail in the section on untested patients below)

This procedure generates four functions (one lasso and one tree-based, for each outcome), which we apply 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 $Y_i^T = 1|T_i = 1$ 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 holdout 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.), we do so 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 holdout 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.

As a basic check of face validity, we can also compare model predictions to another source of predictions: doctors' testing decisions. Overall, our $\hat{y}$ seems to correlate with doctors' decisions on whom to test. We return to this issue in more detail below, but as a simple summary statistic, a logit of testing T_i on model-predicted risk $\hat{y}_i$ yields a coefficient of 0.596 (standard error: 0.006) – doctors are over 8.59 times more likely to test patients in the highest ventile of model predicted risk than the lowest.

3.2 Yield of Testing among Tested Patients

We now turn to the relationship between model predictions and economically meaningful outcomes, starting with the yield of testing. For each Medicare patient tested for heart attack in the hold-out set, we have both an individualized $\hat{y}_i$, the algorithm-predicted individual level risk, and the realized yield of testing. We can use this to explore the testing margin used by doctors. Table 3 shows first the relationship between testing yield and predicted risk. A first observation is that the model discriminates well between high- and low-risk patients, even among the group of patients doctors suspect enough to test: yield is approximately monotonic in predicted risk, and the model is able to identify large groups of patients with very different risk relative to the average rate of revascularization among the tested (13.8%). Indeed, the lowest decile of tested patients in terms of model-predicted risk had only a 2.2% revascularization rate, less than a quarter of the average rate.

Yield units, while clinically relevant, do not necessarily capture the policy implications of these decisions. Luckily, there is a substantial literature on the cost effectiveness of health interventions we can draw on to turn these numbers into more intuitive units of cost per quality adjusted life years. Given the importance of health expenditures for national policy making, there are also objective thresholds for judging a decision cost effective or not, while these vary by country: $50,000 is widely cited in the UK, vs. $100-150,000 in the US (Neumann, Cohen, and Weinstein 2014). We draw on this same literature to perform a simple cost effectiveness calculation for the tests in our sample (with a fuller accounting of individual costs, benefits, and assumptions used from the literature in the Supplement).

We first model the doctor's decision making process as follows.

1. She estimates the probability that a patient is having a heart attack, $\hat{h}_i$.
2. 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.[17]
3. 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 shows the cost effectiveness of tests in units of cost per life year. We can see that the lowest-risk patients that doctors nonetheless choose to test are strikingly low-value: for example,

[17] 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.

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 gives us a very different picture of the problem in two ways: first, it paints a stark picture of overuse of testing in this population. At an average cost effectiveness of $135,859 per life year, we might conclude that testing as a whole is barely cost effective, or slightly ineffective relative to the commonly used thresholds of $100-150,000. But since we no longer have to work in averages, we can inspect the lowest-value tests at the margin directly, where value is lower by a factor of nearly 5. This suggests that the conventional approach of measuring moral hazard, by measuring average yield, actually under-estimates the extent of low-value care. Second, rather than policy prescriptions to test less in general, we can identify specific marginal tests we would 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. So rather than exhorting doctors to test less in general, tailored risk predictions could eventually play a role in decision-making at the individual patient level.

Just as the traditional approach of considering average yield obscures the extent of predictably low-value testing in the lowest-risk patients, it also ignores another important point: the very high returns to testing the highest-risk patients. In the top 10% of tests in terms of model-predicted risk, the cost per life year is only $82,621 per life year, well within even stringent 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 among the tested, we can just as easily generate predictions for the untested. We then look at patients whose predicted risk levels would make them highly cost effective to test, and ask: how often are they tested by doctors? As shown in column (3) of Table 3, while doctors are more likely to test higher risk patients, a surprising finding here is that only 10.7% of patients in the highest risk decile (defined 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 on the part of 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.

Our finding, though, is at best suggestive. Any effort to study under-testing must grapple with a basic econometric problem: we do not observe test results for untested patients, and do not know if they would have ultimately 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: also nothing. But in untested patients, what would have happened? Answering this question is difficult because of *private information* observed by doctors but not the statistical model. Certainly, the algorithm uses a rich set of data: in insurance claims, all prior

diagnosis and procedure codes, capturing results of prior testing; and far more in electronic health records, which include 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 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. Without their test results we cannot conclude one or the other.

A rough calculation illustrates the scope of the private information problem. Consider that, among 76,378 tested patients, 9,407 (i.e., 12.3%) ultimately received prompt revascularization interventions. If the 1,477,329 untested patients had the same rate of revascularization as tested patients with the same model-predicted risk, there would be 138,141 (i.e. 9.35%) revascularization interventions in untested patients. This would in turn imply that doctors were currently diagnosing and treating only 6.38% 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 (ST-ACS), and provide 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 (and any other unobserved factors) drives both doctors' decisions and the yield of testing, model predictions in the untested would be inaccurate. To illustrate this, we replicate our predictive model in 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. Using regular expression matching, we identify for two key ECG findings noted by the cardiologist charged with interpreting the study: 'ST-elevation,' a finding very concerning for heart attack, and 'normal ECG,' which cardiologists use to denote the absence of any abnormality on the study. Table 4 shows two main findings.

1. 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%, $p < 0.001$).

2. These decisions correlate to true risk: yield of testing also depends on ECG features, conditional on risk prediction using structured data. Patients with ST-elevation are 2.5 times more likely to receive interventions than those with high-risk ECGs (80.0% vs 31.5%, $p < 0.001$), while those with a normal ECG are 44% less likely (20.9% vs 37.4%, $p = 0.004$), all conditional on our usual risk prediction without ECG data.

Indeed, 52.2% of patients in the highest-risk quintile of predicted risk did not even have an ECG performed. While some of these decisions may represent errors of omission, in the majority of cases

it is likely to indicate that 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 (to ensure reimbursement for the ECG, which does not happen without a formal interpretation – even if it 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 introduce future information for prediction of a past event.

We would ideally like to read ECG features directly from the waveform recording electrical depolarization of the heart muscle, as detected by electrodes placed on the skin surface, rather than relying on the cardiologist's interpretation. Historically, including the waveform, as opposed to features of the interpretation, would have been difficult. But the advent of deep learning models to handle such data means that including ECG data directly is now tractable.

We thus implement a residual neural network, a variant of the standard convolutional neural network used for deep learning, modeled on the architecture described by Rajpurkar et al. 2017 to incorporate ECG waveforms directly into our model. Specifically, we implement a 34-layer convolutional neural network to predict the same outcome we study above, intervention among the tested. The model takes as input a raw ECG signal, which 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. 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. On average, $\hat{y}_i'$ incorporating the ECG waveform is 18.2% lower than without the waveform. Another way to illustrate this is by calculating, for each individual observation, $p(\hat{y}_i < \hat{y}_i') = 0.740$ (SE: 0.006, derived from 1000 bootstrap samples).

This finding, which is just one channel by which private information can distort conclusions from a predictive model, illustrates the scale of the unobservables problem. Certainly, the ECG is an important factor, and including it in a predictive model is useful – where it is available. But given that it is only one of a myriad of factors that doctors observe that are imperfectly measured or simply unobservable, it is clear that different solutions are required for inference in untested patients.

4.2 Clinical Outcomes in Longitudinal Data

To form a solution to the problem of unobservables, we exploit an important fact about the natural history of heart attack: while we do not know test results in the untested, we do know their eventual outcome, thanks to the longitudinal nature of Medicare claims. If individuals had undiagnosed heart attacks, we ought to see those consequences manifest over time. The specific consequences we might see are all too well known: there is a long tradition of clinical research dating well into the modern era of medicine, which documents in sometimes heartbreaking detail the fate of patients with untreated heart attack. This is primarily because there were no effective treatments, meaning that observational studies of untreated heart attack – including, for example, trials comparing home vs. hospital management of heart attack (Mather et al. 1976; Hill, Hampton, and Mitchell 1978)

– were common until the early 1980s. Precisely because these patients were untested (and as we show, undiagnosed) their outcomes are not masked by treatments.

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.[18] 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 excellent agreement with expert judgment after chart review (e.g., Wei et al. 2014).

Rates of these adverse events in the 30 days after visits are shown in Figure 2. Untested patients in the highest-risk decile have strikingly high rates of adverse events: 3.8% return, only to be diagnosed with heart attack or the cardiac arrest that results 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 their initial encounters: at that time, most patients with realized adverse events were sent home from the emergency department (55.1%) instead of hospitalized, implying that doctors under-estimated their risk. And a majority (60.4%) were diagnosed with conditions like acid reflux or simply 'chest pain' that are considered suspicious for missed heart attack by a substantial literature on diagnostic error, e.g., Wilson et al. 2014.

A key question is whether these rates would constitute under-testing? Certainly, relative to the base rate of adverse events (1.7%), the highest-risk patients are far more likely to succumb to heart attack or death. But what does the rate at which we observe these events mean for whether we should have tested these patients? A useful benchmark here comes 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) are all calibrated to minimize the rate of such adverse events in untested or untreated patients. This line of research, which guides routine testing and management decisions in clinics and hospitals and underlies recommendations from professional societies, gives us objective thresholds for levels of risk that would mandate further evaluation. 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 changes in management decisions (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). These results as a whole lead us to conclude that at least some fraction of high-risk untested patients should have been tested, at least by the standards widely used in clinical research and practice.

4.3 Biomarker Data

Of course, while the evidence we find in Medicare claims data is suggestive, it may not be the best indication of adverse events. After all, these data were originally generated for billing purposes. This introduces some inevitable measurement error – some random, but some less innocent, since

[18] 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.

there are many incentives to 'up-code' visits to support increased reimbursement. So it would be ideal to have a more objective way to measure the nature and severity of heart attack. To do so, we return to the EHR dataset: in addition to the ECG, this dataset also contains longitudinal outcomes, including biomarkers – if the patients returns to care in the 30 days after an untested visit, and the doctor decides to measure the biomarker – which can give us a precise indication of the extent of biological damage to the heart. So we assemble these data on the untested patients in our sample.[19] In the remaining visits, we identified biomarker evidence of heart attack using measured values of cardiac troponin (i.e., cTnT, which measures death of heart muscle cells).

Figure 3 shows, by decile of predicted risk in untested patients, the maximum measured troponin over the six months after ED visits. The usual caveats with electronic health records apply: not only would a doctor need to decide to obtain the test, but the test would need to happen in the health system network of the hospital we study. While this is perhaps less of a concern for patients whose visit led directly to testing —- since positive tests are highly unlikely to lead to discharge from the hospital and potential loss to follow up -— it becomes important for longer-run outcomes in untested patients. So these numbers should be viewed as providing a useful lower bound on the frequency of these outcomes. That said, the results are striking. Among patients in the highest risk decile, a full 18.9% have biomarkers consistent with heart attack at 30 days, and over a third of these have substantial elevations (i.e., cTnT ≥ 0.1); these outcomes are vanishingly rare in the lower risk deciles.

4.4 Over- and Under-testing

So far, we have uncovered evidence of both over- and under-testing. Over-testing was straightforward to measure, using standard cost effectiveness calculations and thresholds. Under-testing is more difficult to measure precisely: certainly, the rates of adverse events seem high relative to accepted clinical standards, but how to arrive at a quantitative estimate of its magnitude? We do so by assuming a simple lower-bound for the extent of under-testing. 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 patients who go on to have a realized adverse event in the month after an untested visit. We apply this conservative bound to identify patients who would have been very 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 prediction, 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 4. 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 4 shows the results of the same procedure, but in units of dollar benefits rather than number of tests; this combines the number of life years saved (at the valuation on the x-axis) and the costs of all tests. Importantly, even though this strategy would on net reduce testing, a large fraction of the benefits

[19] As with Medicare claims, we exclude patients whose heart attack was likely known to doctors in the ED: of the 143,180 untested, 4,946 had a positive troponin in the ED or a diagnosis of heart attack. The assumption we made in the Medicare claims analysis – that these patients likely had a reason for which either testing or revascularization procedures were impossible – can now be confirmed, via a hand-review of a sample of charts. Reasons included patient or family preference, known severe coronary disease refractory to treatment, and other reasons.

of reallocation come from increasing testing for the high-risk untested, a fraction which increases as one's valuation of a life year increases. For example, at $150,000 per life year, 42.8% of net benefits 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).

4.5 A New Look at Variations in Care

The simultaneous existence of both over- and under-testing highlights the deficiencies of a model built solely on incentives. Moral hazard focuses on where doctors draw the threshold in the risk distribution – too low, because of incentives – which results in too many low-risk patients tested. But such a model cannot explain why doctors would fail to test high-risk patients: those 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 – certainly more profitable than a negative test. So why would doctors be testing lower-risk patients before higher-risk patients who are both more medically needy and more profitable? If doctors' threshold is set too low, then nearly all high-risk patients should be tested, and the marginal patients should nearly all be low-risk. Nor does this model has anything to say on errors in risk prediction: patients should be approximately well-ranked by risk in the distribution – this is what allows doctors to dip down into the lower-risk marginal patients to maximize reimbursements and minimize risk.

Worse yet, the major policy prescription of this model can have perverse consequences. Health economics has documented wide variations in the amount of health care delivered, but little variation in outcomes. If the common explanation for this is that marginal patients are low-value, the common solution is that doctors should simply test less. A common trope in the health policy literature is to exhort high utilizing doctors or hospitals to be more like their low utilizing neighbors (e.g., Liao, Fleisher, and Navathe 2016). And a major way by which efforts to change provider incentives (e.g., Obamacare) operate is to let doctors decide where and when to cut back care (e.g., Loewenstein, Volpp, and Asch 2012).

As above, the ability to look at the testing margin through the lens of tailored, individual-level risk estimates can shed new light on this recommendation. Figure 5 shows the results. We first group all US hospitals into quintiles, based on their empirical testing rate in our sample. We then calculate the testing rate in each bin of predicted risk. This shows a striking result: 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. We find 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.

Of course, this evidence is suggestive but not conclusive. 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 of the unobservables problem: even after adjustment for a variety of factors (e.g., our own risk predictions), we can never be sure that variation in a particular aspect of health care is causally linked to the outcomes. Ideally, to be entirely sure that unobservables are comparable between tested and untested patients, we would like to conduct a randomized trial.

Within hospitals or regions, there are settings in which patients might, for idiosyncratic reasons, be more or less likely to be tested. Consider the following two facts regarding common practices in

cardiac testing:

1. It is expensive to maintain staffing of cardiac testing facilities, leading many hospitals to 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.

2. Patients who present to the ED with concerning symptoms on day t are tested immediately to rule out obvious heart attack. If obvious problems are excluded, the decision is made to proceed with cardiac stress testing or catheterization; but this is typically not done on the same day. This is both since 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 and undergo repeat laboratory testing (troponin), and have stress tests or catheterizations on day $t + 1$ (or later).

This led us to hypothesize 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 using differences that straddle a weekend (Saturday) and weekday (Friday). We further hypothesized that these differences would be most pronounced in hospitals with testing facilities on-site (as opposed to those that must always transfer patients out to be tested), and in patients for whom these hospitals were nearby (to avoid capturing patients who may have sought out, or been told to seek out, these hospitals specifically). Practically, for these analyses, 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. 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 6 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.

This natural randomization allows us to do three useful things. First, we can verify that, at least in this range of testing rates, the model estimates predict realized testing yield in the tested, as well as adverse events in the untested, just as well on the weekend as on the weekday. This is shown in Figure 7.[20] Second, we similarly verify the relationship between yield (among the tested) and adverse event rates (among the untested), which 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 8).

[20]We also calculate yield and adverse event rate in marginal patients specifically, 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.

Finally, we can inspect doctors' testing decisions vs. predicted risk: when doctors reduce testing by 19.8%, where in the risk distribution do the marginal patients come from? The results in Figure 9 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. Figure 9 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 – and where we observe the test results in the dropped patients, so can be highly confident of the counterfactual. 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, with 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. This translates into – at the same level of testing reductions – finding 17.3% more patients needing intervention than doctors, saving 17.7% more life years and generating (at the \$150,000 per life year) a surplus of \$150.7 million in our sample.

5 Cognitive Errors

How should we interpret the fact that doctors seem to be both over- and under-testing? 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.

5.1 Boundedness in Physician Judgments

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 2017; 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 10 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 very 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 the first fact that emerges is that doctors appear to use a far simpler model of risk than the algorithm.

The second fact is that the variables included in this simpler model are weighted proportionally to their observed relationship to true risk – in other words, doctors appear to be generally getting 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 different weightings. But in fact, as shown in Figure 11, 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.

Though these results provide evidence that physicians are using a small number of variables effectively, this by no means implies the absence of specific biases. To better understand these, we regress the testing decision t on the full set of X variables (with a regularization penalty), controlling for predicted risk $\hat{y}$ (as produced by our risk predictor, and with no regularization penalty):

$$\hat{\beta}_\lambda = \arg\min_\beta \|t - (\beta X + \gamma \hat{y})\|_2^2 + \lambda \|\beta\|_1$$

The coefficients of this regression are not meant to be conclusive, but rather to motivate more

specific analyses: the quantitatively large elements of β_λ give us clues of where physicians might be over- or under-weighting true risk. This procedure produces two suggestive places to look, each of which we explore below: (i) age stands out as the largest (positive) coefficient, suggesting a focus on how physicians handle demographics; and (ii) injuries and infections in a patient's history have the largest negative coefficients, suggesting a focus on how physicians handle competing diagnoses.

5.2 Demographic Groups, Representativeness and Stereotyping

In forming their risk assessments, physicians readily observe demographics: age, gender and race, which we represent as a set of demographic cells d comprised of individuals with specific combinations of these variables. To study how effectively physicians make use of these variables, we calculate for each demographic group d their true base rate of risk p_d, using the yield of testing. Physicians have an implicit risk they associate with each group q_d. We are interested in $e_d \frac{q_d}{p_d}$. Theories based on representativeness (Kahneman and Tversky 1972) make a clear prediction about e_d: in judging individual risk, physicians will over-weight the component of risk that arises from demographics. In the model of stereotyping of Bordalo et al. 2016, for example, such a prediction follows clearly because physician judgment is roughly of the form: $q_d = h(\frac{p_d}{\bar{p}})$ where $\bar{p}$ is the average risk in the population and $h()$ is a convex weighting function. The convexity of the weighting function ensures that groups with higher than average risk (more representative) have their risks *over-estimated*. As a result, groups that are more representative of heart attack (i.e. have higher than average risk of heart attack) are over-tested. Conversely, less representative groups are under-tested. Notice that this is not a statement of the level of testing but of the level of *error* in testing.

To test this idea, we need to calculate e_d. The implication of representativeness then is that e_d is positively related to p_d: the rate of over-testing is positively related to the base rate of risk. We effectively measure e_d by estimating a testing equation in which we condition on the true risk prediction and include dummies for the demographic bins defined by all age, race and gender combinations. Since we condition on true risk, the coefficients on these demographic variables reflect over-weighting and measure e_d. We can then uses this regression to calculate the rate at which each demographic group would be tested if they had equal risk. This is shown on the x-axis in Figure 12, where we separate those under age 65 from others, as they are a distinct group in the Medicare population (since their eligibility comes from disability rather than age). On the y-axis we show the actual average risk of that group. The figure starkly illustrates how, even after controlling for risk, riskier groups are still tested more than less risky ones. This figure produces fairly direct evidence for a view based on representativeness or stereotyping, where demographic variables are used to form representativeness judgments for heart attack risk.

5.3 Competing Diagnoses and Occam's Eraser

Our second exploration of bias is motivated by understanding how physicians handle competing risks. In deciding whether a patient is having heart attack, physicians also implicitly are making decisions about a range of other, competing diagnoses. In our analysis, we focus on pneumonia since it is a common infection, which was one of the categories of variables that physicians seem to over-weight (negatively) when deciding on whom to test. Most importantly, some of its symptoms overlap quite a bit with heart attack, such that a patient with chest pain or shortness of breath presents a diagnostic dilemma: is this heart attack, or recurrence of pneumonia? Of course the treatments for these two are quite different, making it important to get the diagnosis right.

To understand how physicians treat such a competing risks, we segment patients in the holdout set into three groups: those who had a recent diagnosis (pneumonia in the 30 days before their ED visits), those who contracted pneumonia at a more remote point in time (anytime from one month to three years before their visit), and all others. We then observe the rate at which doctors choose to test these patients, and also the rate at which they experience adverse events.

Figure 13 shows that patients with both recent and remote histories of pneumonia are less likely to be tested. This is not necessarily an error: pneumonia is indeed a credible alternative explanation, and if physicians were able to distinguish these perfectly we would see exactly this. One clue that this might be an error is that the reduction in testing is seen equally across the risk distribution – one might have expected that pneumonia is a less compelling alternative when base rate risk of heart attack is very high. To fully understand whether the reduction in testing is erroneous, we look at those who are untested. Here we see more compelling evidence of a mistake: when physicians choose not to test someone with a history of pneumonia, they have much higher rates of adverse events. So a plausible alternative explanation leads physicians to over-rely on it. Both heart attack and pneumonia – whether a recurrence, or simply persistence of symptoms from a recent bout – could be present in the set of diagnoses that the doctor considers. However, it seems that, doctors excessively hone in on this alternative explanation, omitting the possibility that both can happen.

These results suggest a cognitive error which we call 'Occam's eraser.' Occam's razor reflects an efficient favoring of parsimony. Occam's eraser, on the other hand, suggests an excessive focus on parsimony – the tendency to under-weight a possibility just because an alternative fits the facts. This finding is related to several findings 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).

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. 2015; 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. 2015) 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. 2017): 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. 3730–64.

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.

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

Balogh, Erin P. et al. (2016). *Improving Diagnosis in Health Care*. National Academies Press.

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 (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. (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. (2015). "What does a deductible do? the impact of cost-sharing on health care prices, quantities, and spending dynamics". In: *National Bureau of Economic Research Working Paper*.

Brown, Michael D. et al. (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 (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. (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.

Chetty, Raj, Adam Looney, and Kory Kroft (2009). "Salience and taxation: Theory and evidence". In: *American economic review* 99.4, pp. 1145–77.

Cobb, Leonard A. and W. Douglas Weaver (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 (1989). "Clinical versus actuarial judgment". In: *Science (New York, N.Y.)* 243.4899, pp. 1668–1674.

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

Foy AJ et al. (2015). "Comparative effectiveness of diagnostic testing strategies in emergency department patients with chest pain: An analysis of downstream testing, interventions, and outcomes". In: *JAMA Internal Medicine* 175.3, pp. 428–436.

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

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

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

Ghassemi, Marzyeh 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 (2005). "Diagnostic Error in Internal Medicine". In: *Archives of Internal Medicine* 165.13, pp. 1493–1499.

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. (2017). "National Health Care Spending In 2016: Spending And Enrollment Growth Slow After Initial Coverage Expansions". In: *Health Affairs* 37.1, pp. 150–160.

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

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

Jolly, Sanjit S. et al. (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.

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

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

Kleinberg, Jon et al. (2017). "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. (1999). "Weekend and Holiday Exercise Testing in Patients with Chest Pain". In: *Journal of General Internal Medicine* 14.1, pp. 10–14.

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 (2016). "Increasing the Value of Social Comparisons of Physician Performance Using Norms". In: *JAMA* 316.11, pp. 1151–1152.

Litt, Harold I. et al. (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 (2012). "Incentives in Health: Different Prescriptions for Physicians and Patients". In: *JAMA* 307.13, pp. 1375–1376.

Mahoney, Elizabeth M. et al. (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. (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. (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. (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: *NBER working paper*.

Mullainathan, Sendhil and Ziad Obermeyer (2017). "Does Machine Learning Automate Moral Hazard and Error?" In: *American Economic Review* 107.5, pp. 1–5.

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

Murphy, Sherry L. et al. (2018). "Mortality in the United States, 2017". In:

Neumann, Peter J., Joshua T. Cohen, and Milton C. Weinstein (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 and Thomas H. Lee (2017). "Lost in thought—the limits of the human mind and the future of medicine". In: *New England Journal of Medicine* 377.13, pp. 1209–1211.

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

Paxton, Chris, Alexandru Niculescu-Mizil, and Suchi Saria (2013). "Developing predictive models using electronic medical records: challenges and pitfalls". In: *AMIA Annual Symposium Proceedings*. Vol. 2013. American Medical Informatics Association, p. 1109.

Poldervaart, J. M. et al. (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". In: *Arch Intern Med* 172.19, pp. 1506–1509.

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

Rajkomar, Alvin 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, pp. 436–436.

Redelmeier, Donald A. et al. (2001). "Problems for clinical judgement: introducing cognitive psychology as one more basic science". In: *CMAJ: Canadian Medical Association Journal* 164.3, pp. 358–360.

Rich, Michael W. and Charles A. Crecelius (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. (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). "Principles of categorization". In: *Concepts: core readings* 189.

Roth, Gregory A. et al. (2018). "Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 392.10159, pp. 1736–1788.

Rozanski, Alan et al. (2013). "Temporal Trends in the Frequency of Inducible Myocardial Ischemia During Cardiac Stress Testing1991 to 2009". In: *Journal of the American College of Cardiology* 61.10, pp. 1054–1065.

Schiff, Gordon D. et al. (2009). "Diagnostic Error in Medicine: Analysis of 583 Physician-Reported Errors". In: *Archives of Internal Medicine* 169.20, pp. 1881–1887.

Schor S et al. (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, p. 618.

Schwartz, Aaron L. et al. (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 (2018). *HEART Score Improves ED Care for Low-Risk Chest Pain*.

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, p. 73.

Shreibati, Jacqueline Baras, Laurence C. Baker, and Mark A. Hlatky (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.

Singh, Hardeep et al. (2013). "Types and origins of diagnostic errors in primary care settings". In: *JAMA internal medicine* 173.6, pp. 418–425.

Swap CJ and Nagurney JT (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 (2017). "Attention variation and welfare: theory and evidence from a tax salience experiment". In: *The Review of Economic Studies* 85.4, pp. 2462–2496.

Thaler, Richard (1987). "The psychology of choice and the assumptions of economics". In: *Laboratory experimentation in economics: Six points of view*, pp. 99–130.

Than, Martin et al. (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. (2013). "What is an acceptable risk of major adverse cardiac event in chest pain patients soon after discharge from the Emergency Department?: A clinical survey". In: *International Journal of Cardiology* 166.3, pp. 752–754.

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

Wei, Wei-Qi et al. (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. (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. (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), as well as cost per quality-adjusted life year (2), by ventile of model-predicted risk ($\hat{y}$ in a hold-out set of tested patients. We also show the probability of testing within each ventile of risk, formed in the entire hold-out set, i.e., tested and untested patients alike.

Risk	Testing Rate			Yield		
Quartile	ECG Abnormal	ECG Normal	p	ECG Abnormal	ECG Normal	p
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, and by findings on electrocardiograms (ECGs) done in the ED. 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.

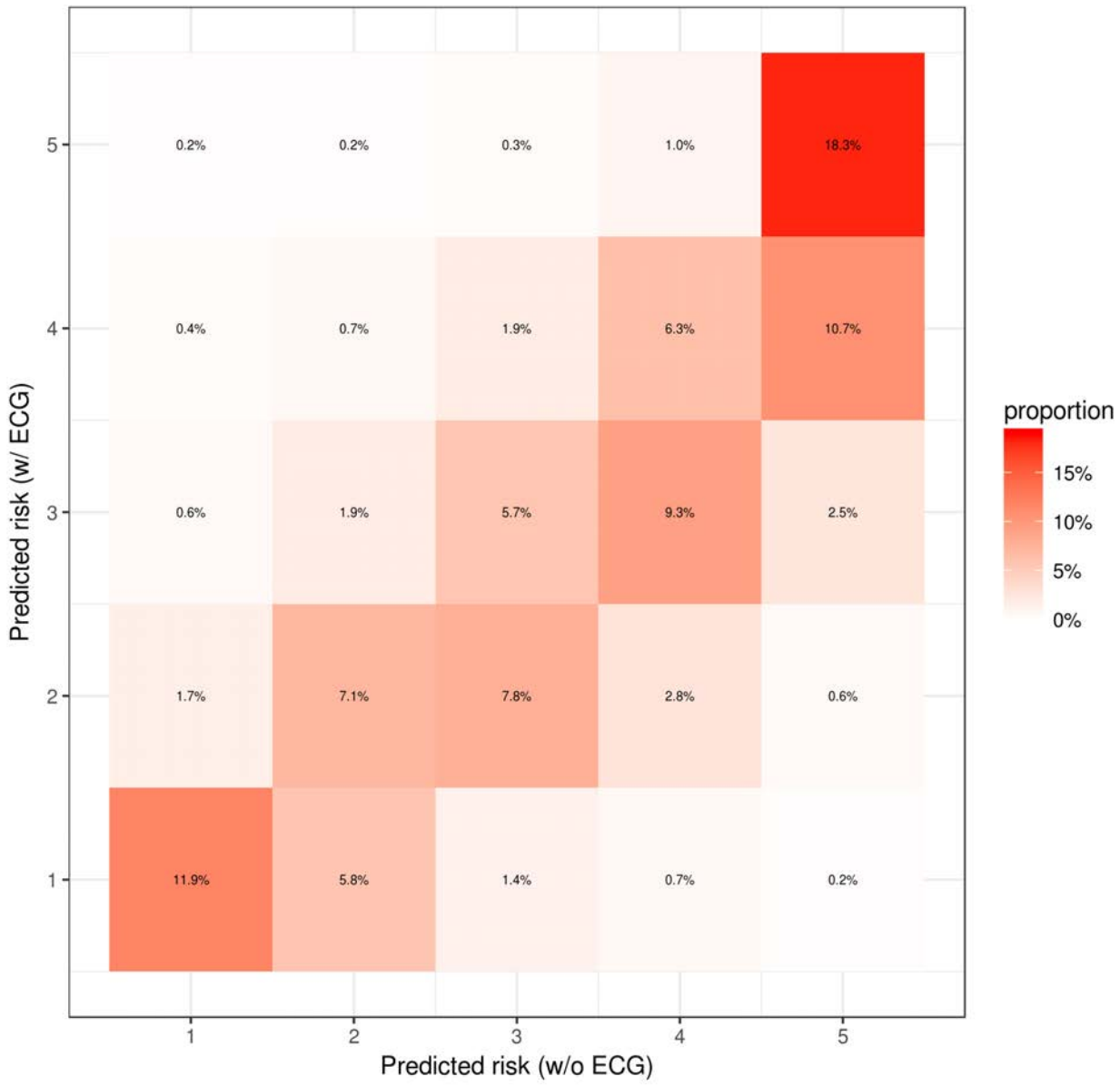

Figure 1: Comparison of model-based risk estimates, with (y-axis) and without (x-axis) incorporation of learned ECG waveform features. Bins are constructed in absolute $\hat{y}$ space on the original predictor.

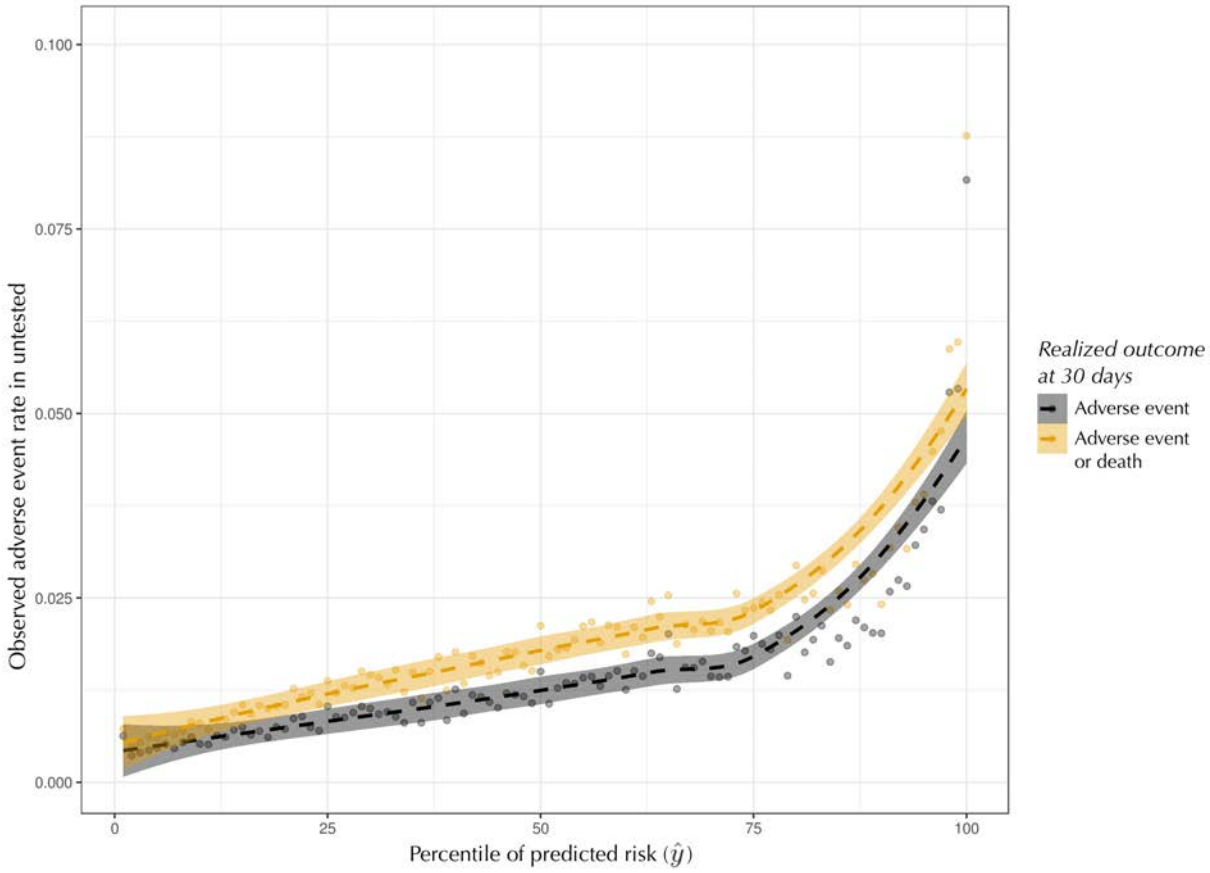

Figure 2: Rates of adverse events (return for heart attack or revascularization, cardiac arrest) and death, by percentile of model-predicted risk ($\hat{y}$) in a hold-out set of untested patients.

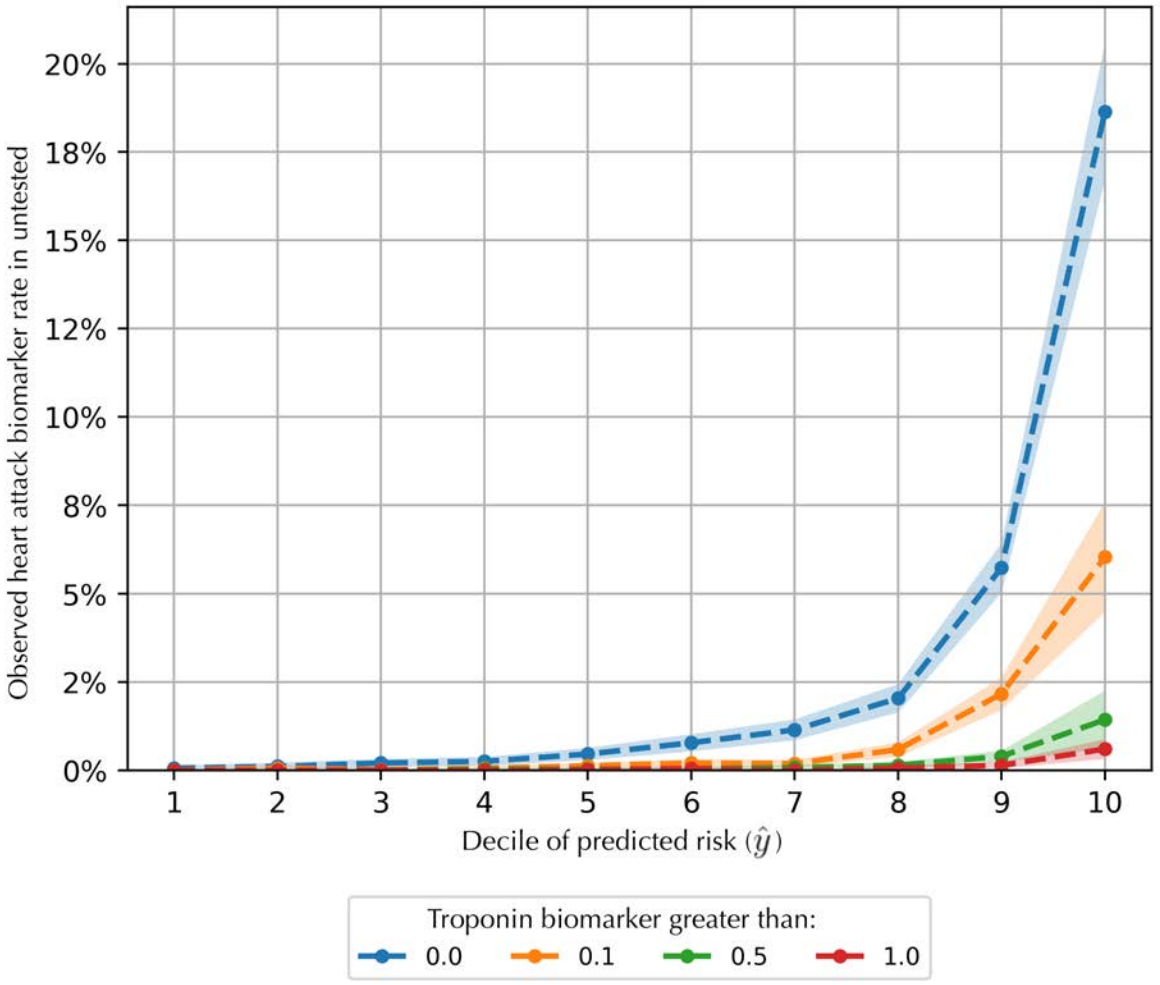

Figure 3: Rates of biomarker-confirmed heart attack, measured by cardiac troponin level in the 30 days after visits, by decile of model-predicted risk ($\hat{y}$ in a hold-out set of untested patients.

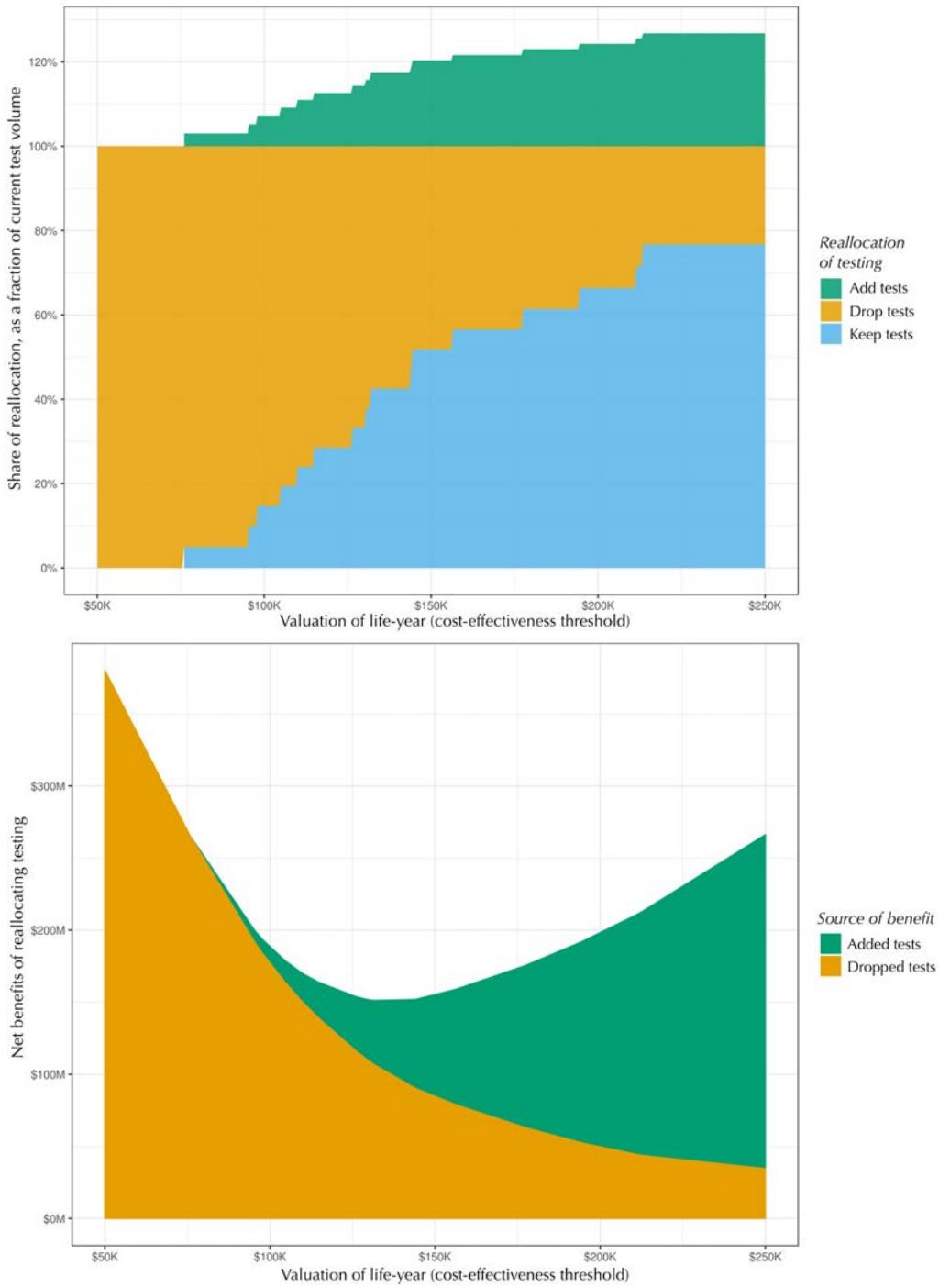

Figure 4: Net over- vs. under-testing at specific cost effectiveness thresholds. Top panel: Fraction of tests currently done that we would keep or drop; and how many of the untested we would choose to test (using a lower bound based on realized heart attack). Bottom panel: Net benefits of reallocating 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. Figures in the text convert these numbers to annual rates in the Medicare population as a whole (though with all sample restrictions preserved).

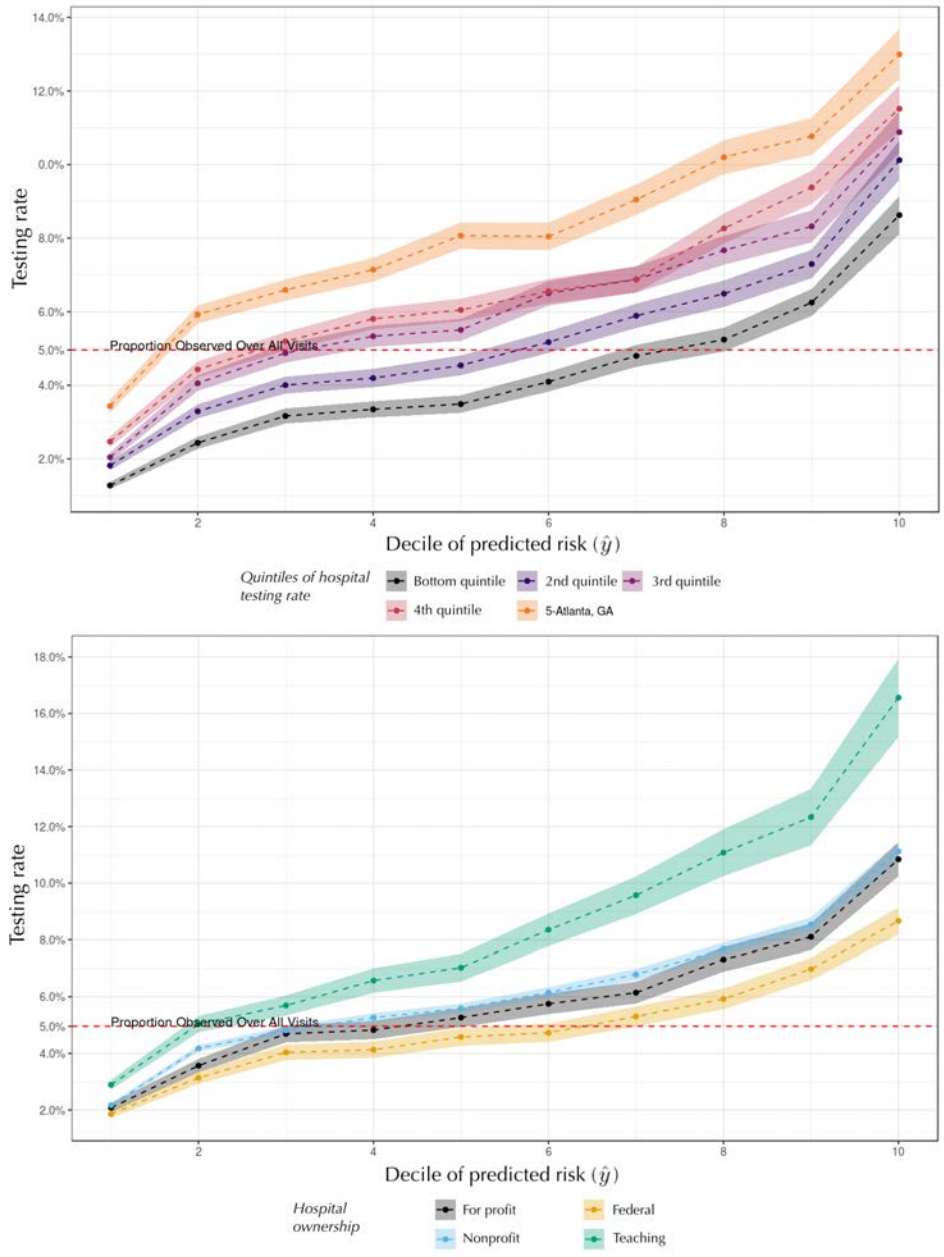

Figure 5: Difference in testing rates across hospitals, by distribution of predicted yield $\hat{y}_i$. Top panel: by quintiles of testing rate and labeled by largest metropolitan area. Bottom panel: by ownership structure. Hospitals with unknown ownership, and those missing Dartmouth data on HRR and spending are excluded.

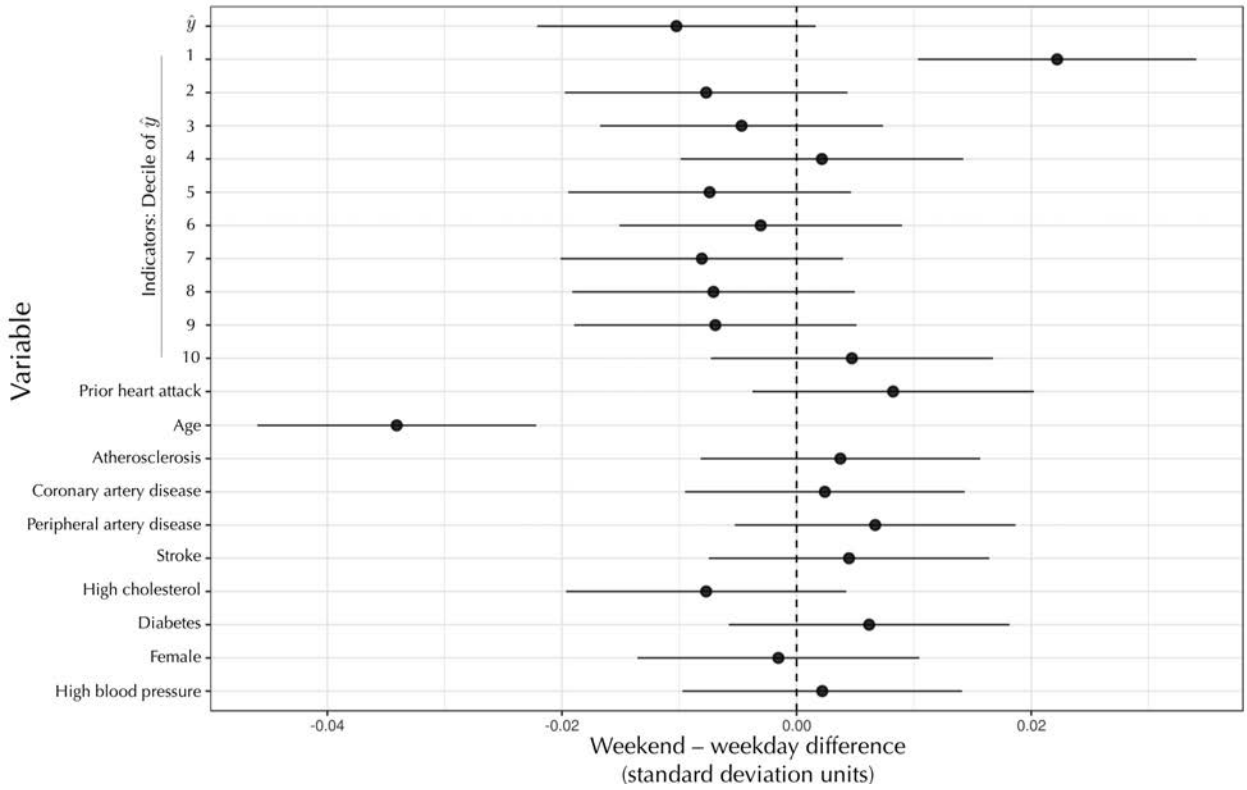

Figure 6: Characteristics of patients visiting on weekend vs. weekdays: Balance on $\hat{y}_i$ (first row), indicators for decile of $\hat{y}_i$ (rows 2-11), and other relevant observables. All differences are conditional on geography (i.e., hospital referral region) and year.

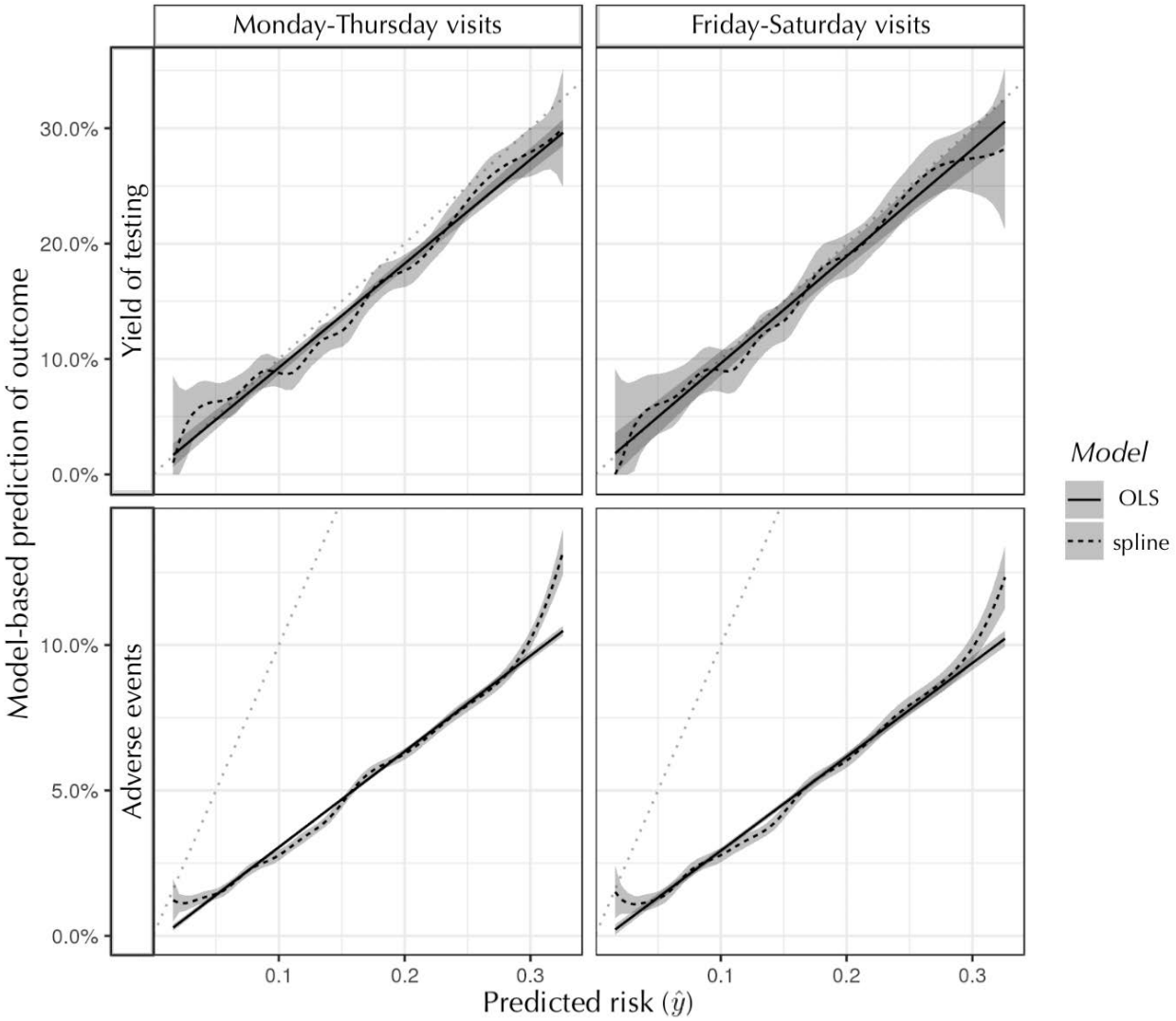

Figure 7: Realized yield of testing vs. decile of predicted yield $\hat{y}_i$ in the tested, weekend vs. weekday. Because the samples are small, we fit this using (1) OLS of the outcome (yield or adverse event) on an indicator for the weekend, $\hat{y}_i$, and an interaction term; and (2) a spline version 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.

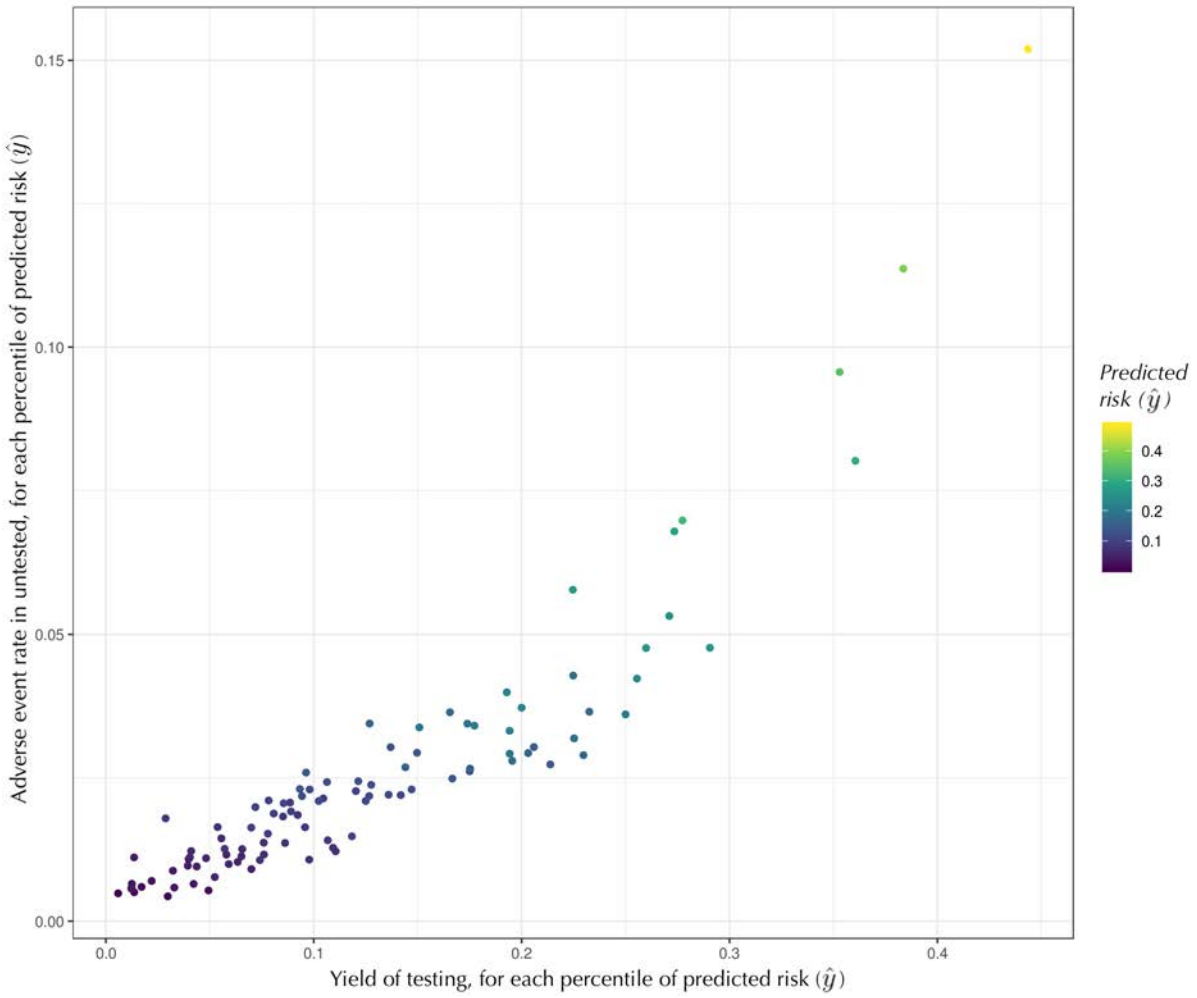

Figure 8: Yield in the tested vs. adverse event rate in the untested, by percentile bin of $\hat{y}_i$ in the weekend vs. weekday sample.

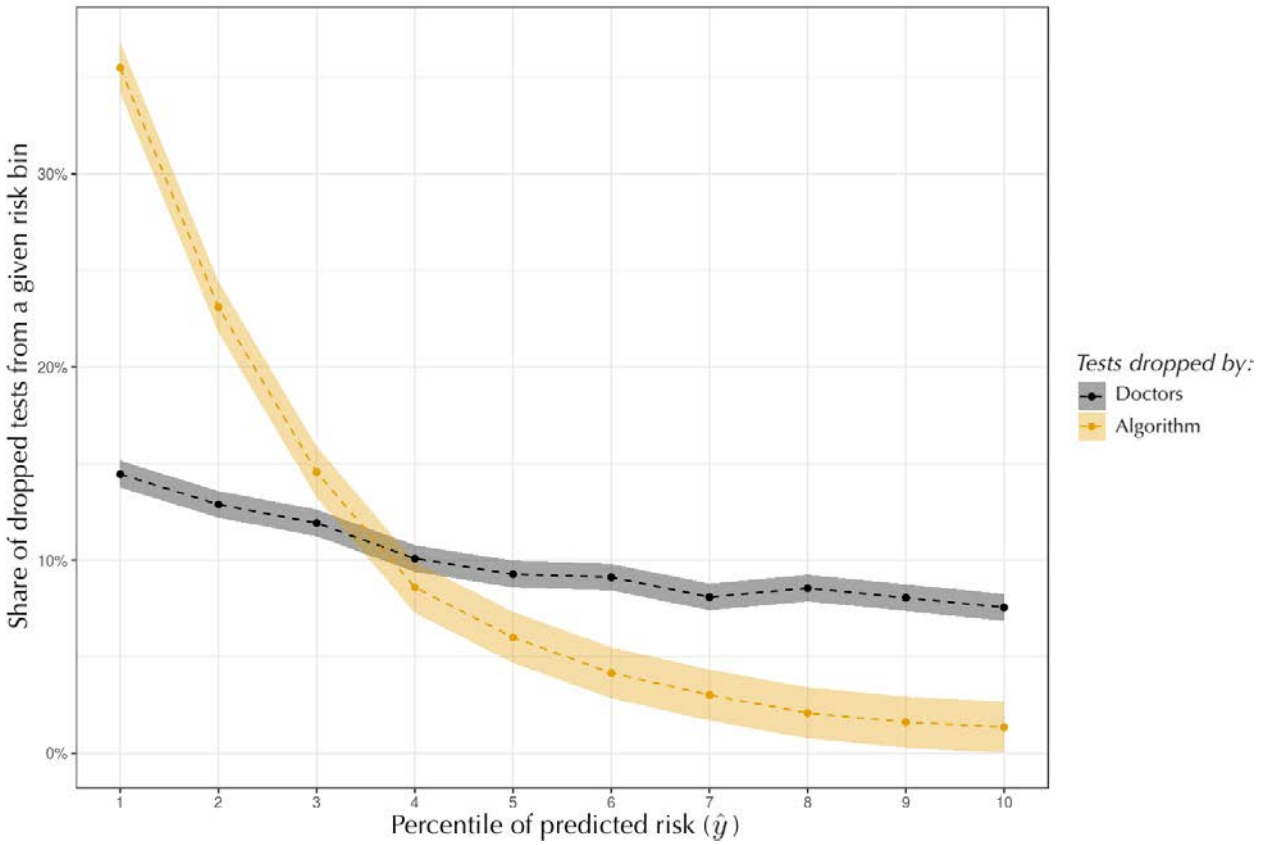

Figure 9: Risk distribution of marginal patients: Difference in testing rates, weekday vs. weekend, by distribution of predicted yield $\hat{y}_i$, comparing observed doctor testing decisions vs. simulated algorithm 'decisions' (conditional on geography and year).

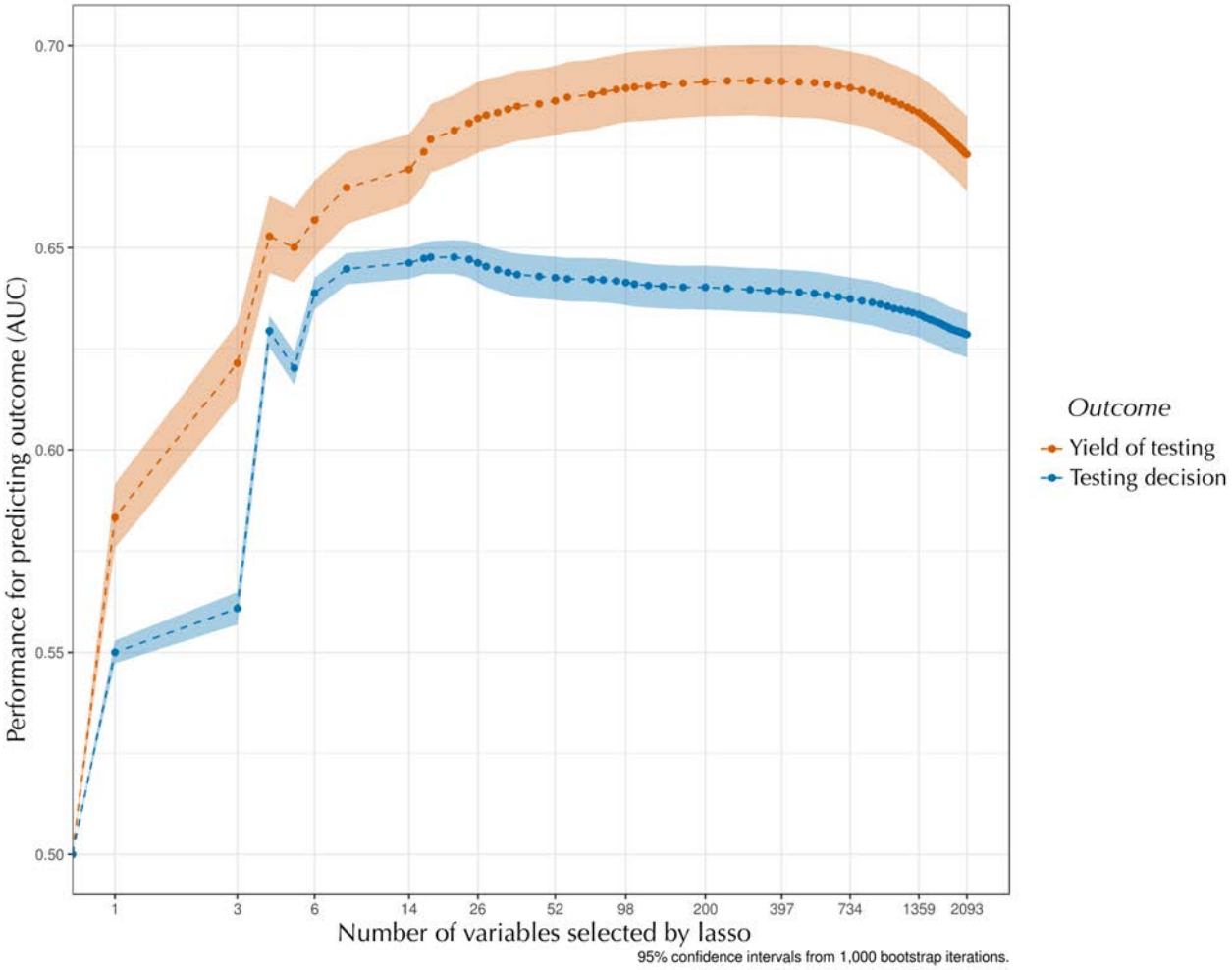

Figure 10: Predictive performance of simplified models of true risk. Models from the lasso path (x-axis: number of non-zero variables in the model) are used to predict true risk (top line) and doctors' testing decisions (bottom line). Performance is measured by area under the curve (AUC).

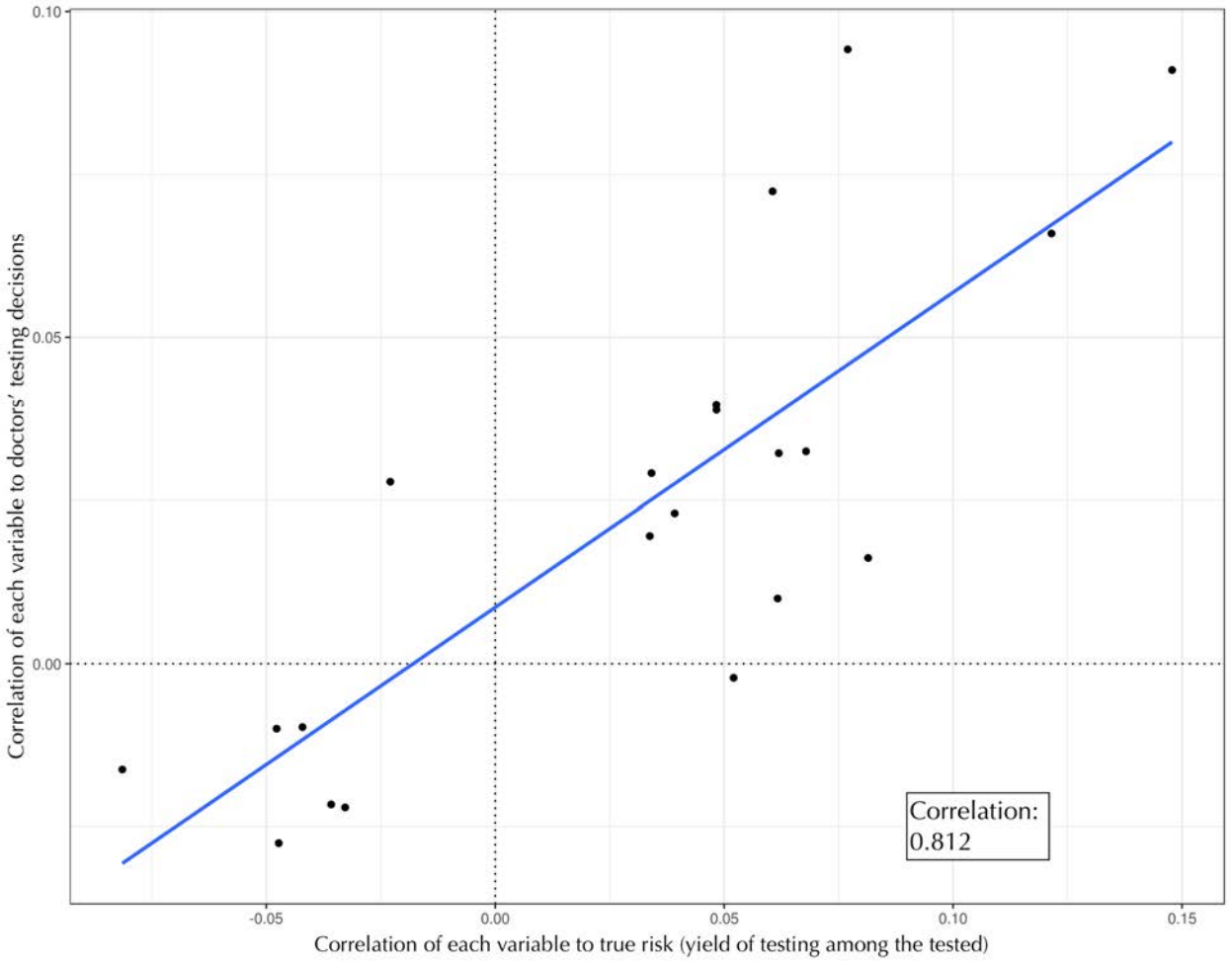

Figure 11: For the 21 variables included in the model that best predicts doctors' testing decisions, univariate correlations of each X with the testing decision (y-axis) and true risk (x-axis).

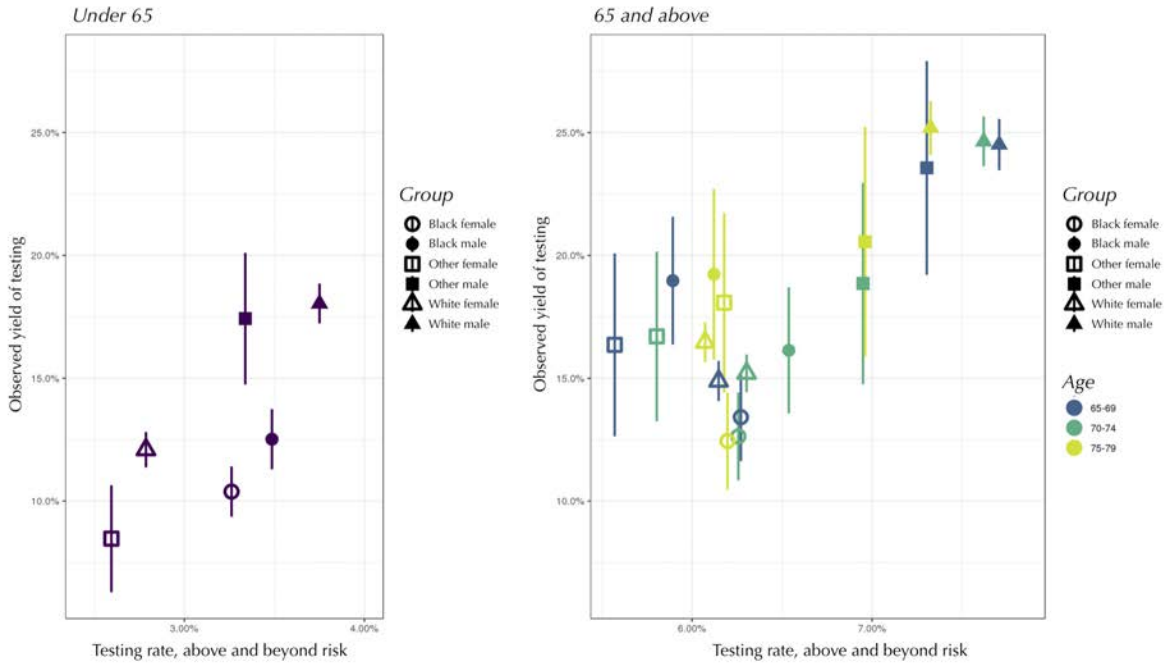

Figure 12: Relation between testing errors and demographic group risk: For each demographic groups defined by age, race, and sex cells, we graph their rate of testing errors (x-axis) against the baseline risk of that group (y-axis). Groups that are high on the x-axis are those that are more likely to be tested above and beyond their risk.

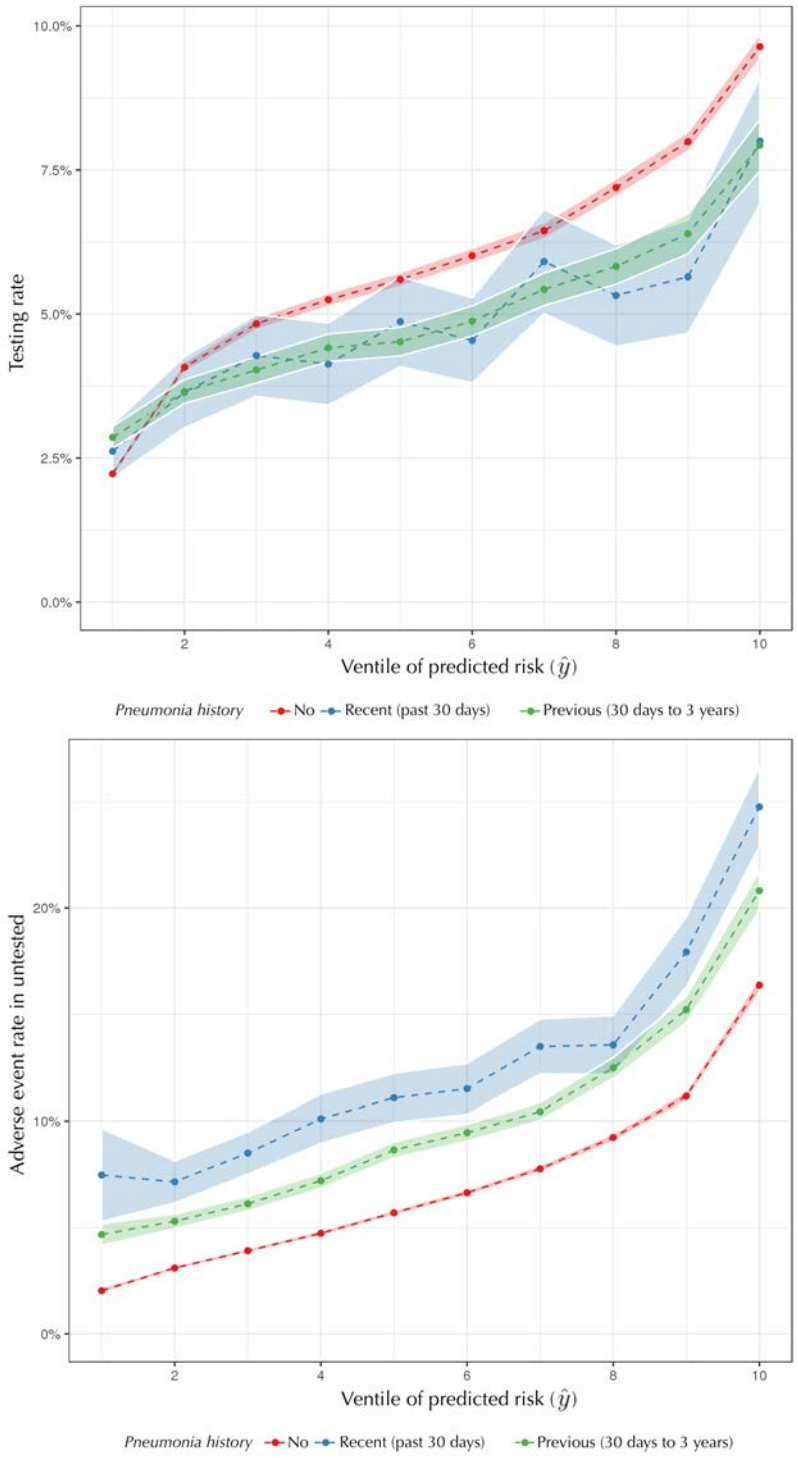

Figure 13: Difference in testing rates and rates of adverse events (6 months after visits) for patients with recent, remote, and no diagnosis of pneumonia, by distribution of predicted yield. $\hat{y}_i$.