THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. 137 2022 Issue 2

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

SENDHIL MULLAINATHAN AND ZIAD OBERMEYER

We use machine learning as a tool to study decision making, focusing specifically on how physicians diagnose heart attack. An algorithmic model of a patient's probability of heart attack allows us to identify cases where physicians' testing decisions deviate from predicted risk. We then use actual health outcomes to evaluate whether those deviations represent mistakes or physicians' superior knowledge. This approach reveals two inefficiencies. Physicians overtest: predictably low-risk patients are tested, but do not benefit. At the same time, physicians undertest: predictably high-risk patients are left untested, and then go on to suffer adverse health events including death. A natural experiment using shift-to-shift testing variation confirms these findings. Simultaneous over- and undertesting cannot easily be explained by incentives alone, and instead point to systematic errors in judgment. We provide suggestive evidence on the psychology underlying these errors. First, physicians use too simple a model of risk. Second, they overweight factors that are salient or representative of heart attack, such as chest pain. We argue health care models must incorporate physician error, and illustrate how policies focused solely on incentive problems can produce large inefficiencies. JEL Codes: C55, D8, D84, D9, I1, I13.

*Authors are listed in alphabetical order. We acknowledge grants from the NIH (DP5OD012161, P01AG005842) and the Pershing Square Fund for Research on the Foundations of Human Behavior. We thank Amitabh Chandra, Ben Handel, Larry Katz, Danny Kahneman, Jon Kolstad, Andrei Shleifer, Richard Thaler, and five anonymous referees for thoughtful comments. We are deeply grateful to Cassidy Shubatt, as well as Adam Baybutt, Shreyas Lakhtakia, Katie Lin, and Advik Shreekumar, for research assistance.

© The Author(s) 2021. Published by Oxford University Press on behalf of the President and Fellows of Harvard College. All rights reserved. For Permissions, please email: journals.permissions@oup.com
The Quarterly Journal of Economics (2022), 679–727. https://doi.org/10.1093/qje/qjab046.
Advance Access publication on December 3, 2021.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

I. Introduction

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

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

This approach reveals two kinds of allocative inefficiency in how physicians test. First, many patients who predictably will not benefit from testing are nevertheless tested. We quantify the value of a test here using the treatment benefits it produces (allowing for the fact that the test itself is imperfect), expressed in cost per life-year saved. By this measure, 62% of tests cost more than $150,000 per life-year. Algorithmic predictions are crucial in uncovering these low-yield marginal tests. Had we instead followed the usual approach of using overall average yields to assess efficiency, we

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

would have concluded that testing as a whole is cost-effective, at $89,714 per life-year (Weinstein et al. 1996; Sanders et al. 2016). Machine learning is useful for capturing such patient-level heterogeneity.

Second, at the same time, many patients who predictably would benefit from testing nevertheless go untested. One hint of this problem, resembling Abaluck et al. (2016)'s earlier work, is that physician choices deviate from a structural risk model: we also find that physicians fail to test many apparently high-risk patients. By themselves, though, such deviations do not establish error, as we do not know what the test results would have been. Physicians may have valid reasons for leaving these patients untested, some of which may be unobserved in our data (and thus to the algorithm): how the patient looks, what they say, the results of X-rays or electrocardiograms (ECGs). The problem cannot be solved by imputing outcomes to the untested.[2]

Health outcomes in the untested provide a way to empirically assess these choices. In the 30 days after their visit, high-risk untested (and thus untreated) patients exhibit the well-known signs of missed heart attack: "major adverse cardiac events" at rates well above existing clinical guideline thresholds for heart attack.[3] One-third of these events lead to death. So these patients appear to have indeed been at high risk. Still, it is possible that physicians recognize this risk but choose not to test because they deem patients unsuitable for invasive treatments. We find evidence to the contrary. For example, a large fraction do not even receive an ECG or other very low-cost, noninvasive tests given to any patient with even a small suspicion of heart issues. Physicians simply seem to overlook the risk for these patients.

2. We illustrate using ECGs, typically missing from research data sets and effectively an unobserved variable to our algorithm: we only have them for a subset of patients (and do not use them in the main analyses). But for this subset, incorporating waveform data via deep learning decreases predicted risk for 97.5% of patients, and 100% of the highest-risk untested, suggesting that predictions are confounded for the untested. Despite growing attention to the selective labels problem, similar biases pervade much of machine learning (Kleinberg et al. 2018; Kallus and Zhou 2018; Rambachan 2021).

3. Such decision rules (e.g., TIMI, GRACE, HEART) are commonly implemented in emergency medicine. We do not take a stance on whether they are physiologically optimal, only that they represent current physician understanding of who should be tested. If physicians use private information in deciding not to test apparently high-risk patients, adverse-event rates should be low.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

For more direct evidence of undertesting, we rely on a natural experiment: a patient's arrival time determines which staff members see them, and staff vary in their tendency to test for heart attack. Conditioning on the visit's hour and day, this provides plausibly exogenous shift-to-shift variation in testing rates.[4] We find that higher-testing shifts do not show statistically significant effects on health outcomes on average, indicating so-called "flat of the curve" health care: more testing yields little return (Fisher et al. 2003). But as before, averages obscure heterogeneity. Predicted high-risk patients benefit from more testing: in the subsequent year, those who arrive during the highest-testing shifts have significantly lower mortality (2.5 percentage points, or 32%), making these additional tests highly valuable.[5] Undertesting is also quantitatively important: we simulate a range of policy counterfactuals that put the size of the undertested set between 15.6% and 99.5% of the currently tested set.

Why do physicians both over- and undertest? Comparing physician decisions to algorithmic predictions suggests several sources of error. We first find evidence of bounded rationality: limits in cognitive resources such as attention, memory, or computation (Simon 1955; Mullainathan 2002; Sims 2003; Gabaix 2014, 2019; Bordalo, Gennaioli, and Shleifer 2020). The risk model that best predicts physician testing is much simpler than the one that best predicts true test outcomes. By way of analogy, physicians seem to overregularize (Camerer 2019). We also find evidence that physicians overweight salient risks (Tversky and Kahneman 1974; Bordalo, Gennaioli, and Shleifer 2012), such as those due to demographics and symptoms. Finally, they overweight symptoms that are representative (stereotypical) of heart attack (Kahneman and Tversky 1972; Bordalo et al. 2016). For example, patients with chest pain, a salient and representative symptom, are particularly overtested.

Health care models have long emphasized moral hazard: paying for tests, rather than outcomes, results in too much testing

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

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

(Arrow 1963; Pauly 1968). Recent work has broadened this perspective to include skill differences, comparative advantage, and error as sources of inefficiency (Abaluck et al. 2016; Chandra and Staiger 2020; Chan, Gentzkow, and Yu 2022).[6] We extend this literature by providing evidence of substantial undertesting, methodologically showing an important role for machine learning, and by uncovering some potential sources of error.

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

II. Context and Framework

II.A. Medical Context

The coronary arteries provide blood flow to the heart, allowing it to pump. A blockage in those arteries abruptly reduces blood flow and kills a patch of heart muscle, an event called an acute coronary syndrome (ACS).[7] Its consequences can be immediate (e.g., arrhythmia, sudden death) and longer-term (e.g., fatigue, heart failure). Randomized control trials have shown two treatments greatly improve mortality and morbidity if delivered promptly: inserting a flexible metal tube into the blocked artery to restore flow (stenting), and for severe cases, bypassing the

6. Abaluck et al. (2016) highlight how errors may produce both under- and overtesting. Chan, Gentzkow, and Yu (2022) show how differences in skill alone, without incentives, can produce what appears to be overtesting. Chandra and Staiger (2020) focus on comparative advantage: because some health systems specialize and focus on certain tests and conditions, they may appear to overtreat those. There is also a large clinical literature on error and its behavioral sources (Dawes, Faust, and Meehl 1989; Elstein 1999; Redelmeier et al. 2001; Ægisdóttir et al. 2006).

7. This is colloquially called a heart attack. We use "blockage" to refer to ACS, to distinguish it from a broader category of problems involving damage to the heart from any cause.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

blockage through open-heart surgery.[8] Timely treatment, though, requires timely diagnosis, a challenging task in the emergency department (ED). Even life-threatening blockages have subtle symptoms, for example, a mild squeezing in the chest, shortness of breath, nausea, or weakness—symptoms that also arise from more benign conditions, such as acid reflux, viral infections, and muscle strain. Any suspicion of blockage triggers two simple, noninvasive tests: first the ECG, which measures the electrical activity of the beating heart and can diagnose acute disturbances. Second, a laboratory test called troponin, a component of heart muscle that, when detected in the bloodstream, implies the death of heart muscle cells. Both help estimate the likelihood of blockage and the urgency of the problem. But no test done in the ED can actually diagnose a blockage.

The definitive test for blockage is cardiac catheterization, an invasive procedure carried out in a dedicated laboratory, separate from the ED. A cardiologist inserts an instrument into the coronary arteries, injects dye, and visualizes the presence and location of blockages via X-ray. If a blockage is found, a stent is inserted to open it during the same procedure. An alternative testing pathway adds a step before catheterization: stress testing. This increases patients' heart activity (e.g., by exercising on a treadmill or with a drug). If supply is limited by a blockage, this excess demand will be detected via heart monitoring. The advantage of stress tests is that they are less expensive and noninvasive: if negative, an invasive catheterization can be avoided. The disadvantage is that if positive, the patient still needs catheterization to deliver the stent, and precious time has been wasted.

The proliferation of both tests has been part of the dramatic reductions in rates of missed blockages in the ED. Before widespread testing, miss rates were substantial: between 2% and 11% of blockages went undiagnosed in the ED (see Pope et al. 2000). Both tests, though, are costly: thousands of dollars for stress tests and tens of thousands for catheterization, plus overnight observation and monitoring before testing. They also have health risks, particularly catheterization, which is invasive. In addition to a large dose of radiation, it involves injection of

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

dye that can cause kidney failure, a risk of arterial damage, and stroke (Hamon et al. 2008). The decision to test must weigh potential treatment gains against these costs.

II.B. Framework

In our model, patients are characterized by a feature vector (X, Z) and drawn from a fixed distribution over (X, Z). Both X and Z are observed by the physician, but only X is recorded in the data. A blockage $B = 1$ occurs with probability $b(X, Z)$, and a test T for blockage yields a positive outcome with $Pr(Y = 1|B, X, Z) = Pr(Y = 1|B) = p + B(q - p)$, where p and q are the false and true positive rate, respectively, and $q > p$. Stenting S can treat the blockage, but because the procedure requires knowing where to place the stent, we assume it can only be done on patients with a positive test $Y = 1$.[9] Moreover, medical ethics would make treatment without testing dubious. (More details are in Online Appendix A.1.C.) B, T, Y, and S are all binary variables, and testing and stenting cost c_T and c_S, respectively.

Using potential outcomes notation, a patient's health is a continuous variable W^S whose value depends on stenting:

$$W^S = W - B(\eta - S\tau^K),$$

with $E[W|X, Z, B] = w(X, Z)$. A blockage harms health by η; stenting partially offsets this harm by τ^K. The binary variable K denotes patients who benefit less from treatment than others (e.g., the frail); they are said to be contraindicated. We assume $\tau^K = \tau - \theta K$, where τ is the baseline treatment effect and the constant θ captures the diminished benefits for contraindicated patients due to their particular health risks from invasive treatment.[10] Patients are contraindicated with probability $k(X, Z)$; physicians know $k(X, Z)$ because it captures known risks, based on current medical knowledge. We define $\tilde{\tau}^K = E[W^1 - W^0|Y = 1, K]$, the average benefit of treating everyone in the population with a positive test. It differs from the average benefit of treating everyone with a blockage, $\tau^K = E[W^1 - W^0|B = 1, K]$, because the test has false

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

10. Practically, those with $K = 1$ may also have higher (health) costs of testing itself, but we omit this for simplicity; it does not change our core empirical results, which focus only on the $K = 0$ population.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

positives and negatives. Randomized trials of stenting, because they enroll only those who show no contraindications and test positive, estimate $\tilde{\tau}^0$. Based on those results, we assume that $\tilde{\tau}^0 > 0$. Finally, an untreated blockage can lead to adverse events after the visit, denoted by a binary variable A. Adverse events occur with probability $\mu + B(\zeta - S\phi)$, so that if $\phi > 0$ stenting reduces their occurrence.

For a patient with characteristics (X, Z), socially optimal testing and treatment would maximize expected health ($E[W^S|X, Z]$) net of costs:

$$\max_{S,T} w(X, Z) - b(X, Z)(\eta - S\tau^K) - c_T T - c_S S,$$

subject to the constraint that only tested patients with a positive test can be stented.[11] Physicians, however, may have a different objective. They maximize

$$\max_{S,T} w(X, Z) - h(X, Z)(\eta - S\tau^K) - (c_T - \nu)T - c_S S.$$

Physician objectives deviate from social objectives in two ways. First, they derive additional benefit $\nu > 0$ from testing, for example, they are paid by the test. Second, they may misestimate the probability of a blockage as $h(X, Z)$ rather than $b(X, Z)$. Given these differing objectives, the socially optimal and the physician testing rule differ:

$$\text{Socially optimal testing: Test iff } b(X, Z) > \frac{c_T + pc_S}{q\tau^K - c_S(q - p)},$$

$$\text{Physician testing: Test iff } h(X, Z) > \frac{c_T - \nu + pc_S}{q\tau^K - c_S(q - p)}.$$

Private benefits from testing and misestimation of risk produce clear inefficiencies: ν lowers the threshold for testing, and $h(X, Z)$ distorts who is perceived as above that threshold.[12]

11. Notice in this setup, testing only benefits health by affecting treatment; it has no other indirect health benefits (such as through information generated for later use). We discuss in greater detail how testing affects stenting in Online Appendix A.1.C.

12. These two equations characterize testing. Treatment is more straightforward: both the physician and socially optimal rules treat all patients with a positive test result.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

To empirically test for such distortions, note that any subset of patients defined by (X, Z) is either above or below the threshold for efficient testing. Those above the threshold should always be tested, and their yield rate should be sufficiently high; those below the threshold should never be tested, and they should have few adverse events. To establish inefficiencies, therefore, we only need to find patient pools that are either (i) tested but have low average yield, or (ii) untested but have high adverse-event rates. The following lemma formalizes this logic.

LEMMA 1. Consider any set of patients defined by a set of characteristics $\mathcal{V}$.

Suppose the tested patients in this set have lower than average yield, $E[Y|(X, Z) \in \mathcal{V}, T = 1] < E[Y|T = 1]$, and their testing and yield rates further satisfy:

$$\underbrace{E[T|(X, Z) \in \mathcal{V}] > 0}_{\text{tested}} \text{ and } \underbrace{E[Y|(X, Z) \in \mathcal{V}, T = 1] < \frac{c_T}{\tilde{\tau}^0 - c_S}}_{\text{but low yield}},$$

then $\mathcal{V}$ is called overtested and eliminating all testing in $\mathcal{V}$ increases efficiency.

Suppose instead the testing and adverse-event rates in $\mathcal{V}$ satisfy:

$$\underbrace{E[T|(X, Z) \in \mathcal{V}, K = 0] < 1}_{\text{untested}} \text{ and}$$

$$\underbrace{E[A|(X, Z) \in \mathcal{V}, K = 0, T = 0] > \mu + \zeta\left(\frac{c_T + pc_S}{q\tau^0 - c_S(q - p)}\right)}_{\text{but high adverse events}},$$

then $\mathcal{V}$ is called undertested and testing all $K = 0$ patients in $\mathcal{V}$ increases efficiency.

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

Proof. Consider a set of patients $\mathcal{V}$, and define $\bar{T}_{\mathcal{V}} = E[T|(X, Z) \in \mathcal{V}]$, $\bar{Y}_{\mathcal{V}} = E[Y|(X, Z) \in \mathcal{V}, T = 1]$, and $\bar{A}_{\mathcal{V}} = E[A|(X, Z) \in \mathcal{V}, T = 0, K = 0]$. First, suppose $\mathcal{V}$ satisfies the

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

conditions for being overtested. If we were to stop testing all tested patients in $\mathcal{V}$, we would save $c_T \bar{T}_{\mathcal{V}}$ per test. But we would no longer get the benefits of the resulting treatments. Because the $Y = 1$ patients (and only those) get treated, these gains come from fraction $\bar{T}_{\mathcal{V}}\bar{Y}_{\mathcal{V}}$ of patients. The net benefit of treating these patients is equal to $\bar{T}_{\mathcal{V}}\bar{Y}_{\mathcal{V}}(x\tau^0 - c_S)$ where x is the fraction of these patients that have a blockage. Tests are wasted if this is less than $c_T \bar{T}_{\mathcal{V}}$ or equivalently if $\bar{Y}_{\mathcal{V}} < \frac{c_T}{x\tau^0 - c_S}$. We can upper bound $x\tau^0$ with $\tilde{\tau}^0$, the average benefit of treating all positive patients, because we have assumed that tested patients in $\mathcal{V}$ have lower than average yield; thus they have lower than average rates of blockage. As such, we can say that the tests in $\mathcal{V}$ are wasted if $\bar{Y}_{\mathcal{V}} < \frac{c_T}{\tilde{\tau}^0 - c_S}$, which is true given the definition of overtested.

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

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

$$\text{Test iff } b(X, Z) > \frac{c_T + pc_S}{q\tau^K - c_S(q - p)},$$

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

Several points are worth noting about this lemma. First, it illustrates the role of machine learning in our analysis: it serves to identify candidate subsets $\mathcal{V}$ where inefficiencies might be present. Second, once identified, inefficiencies are evaluated using observed outcomes: there is no imputation of outcomes. Instead, the key calculations rely only on measured quantities: yield Y for the tested and adverse events A for the untested. Similarly, the relevant thresholds are defined using the clinical literature, as we describe in detail below.[13] Third, it allows physicians to have

13. The adverse event threshold in the lemma cannot be easily stated in terms of model primitives (i.e., the risk of blockage, the imperfect performance of testing,

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

access to information Z that the algorithm does not: it holds for subsets $\mathcal{V}$ identified using only X. One crucial bit of information, though, must be treated carefully: to identify undertesting, we must know $K = 0$. To do so, in the empirical work, we initially assume that $k(\cdot)$ depends only on X, but we weaken this assumption in Section IV.C to allow for it to depend on X and Z. Finally, the lemma links the evidence to an underlying model of physician behavior. Moral hazard alone (bad incentives) can produce overbut not undertesting; misprediction, however, can produce both.

It is useful to contrast this model with two others. Chan, Gentzkow, and Yu (2022) model radiologists who receive a noisy signal about patient risk and choose a diagnostic threshold.[14] While superficially analogous to $h(X, Z)$ and ν, a crucial difference is that in their model physicians are aware that their signal is noisy (and compensate for it, e.g., by testing more to reduce their miss rate). Physicians in our model are unaware of their errors and view their predictions as correct. Our model is closest to Abaluck et al. (2016), who also model physician error. The key difference with them is in how we characterize undertesting: we do not assume $b(X) = b(X, Z)$, that is, that the econometrician can recover an accurate model of the risk of blockage with respect to the physician's information, nor define undertesting as deviations of decisions from predicted risk. Instead, we assume measured health outcomes reflect undiagnosed blockage and use these to characterize undertesting.

III. Data and Methods

Our primary data come from the electronic health records (EHRs) of a large urban hospital from January 2010 to May 2015. It is an academic medical center, consistently ranked in the top 10 best in the country and affiliated with a top-ranked medical school, thus widely believed to provide high-quality care. We begin with all visits to the ED in that period, then exclude patients 80 years or older, those with poor prognosis like known metastatic cancer or dementia, those with hospice or nursing home care, those with

the effect of treatment on health) because several key parameters (i.e., p, q, μ, ζ, ϕ) are unknown.

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

a known recent blockage (or treatment of one), and those who died in the ED before they could be sent for testing.[15]

We observe the patient's main symptom, but do not exclude those with apparently obvious noncardiac problems to avoid potentially arbitrary judgments. While some cases seem clear (e.g., an ankle sprain), many are not: blockage can present in various ways. Worse, we do not observe all of a patient's symptoms, only the one judged most important by the triage nurse.[16] Instead, we use the full sample, and include recorded symptoms in our predictor to make it an empirical question. By including cases highly unlikely to be a blockage, the algorithmic prediction task does become harder: very high-risk cases are comingled with (effectively) zero-risk patients. If it fails, it will appear as an inability to separate high-risk patients from less risky ones. Our final sample has 246,265 ED visits (indexed by j), by 129,859 patients (indexed by i).

III.A. Definitions of Key Variables

In this sample, we define testing $T_{ij} = 1$ if patient i has procedure codes for either stress testing or catheterization in the 10-day window (inclusive) following visit j.[17] We define treatment $S_{ij} = 1$ if there is a procedure code for stenting or open-heart surgery (CABG) in the 10-day window following the visit.

To define test yield Y_{ij}, we rely on the principle that a positive test implies stenting: a cardiologist should not subject a patient to the risks of emergency catheterization unless she has already decided the patient would benefit from a stent if a blockage is detected. So we set $Y_{ij} = S_{ij}$ for the tested. As we discuss further in Online Appendix A.1.C, physicians may overtreat conditional on test results (e.g., because of moral hazard or false-positive tests). One might worry this by itself could artificially produce the results we find. It does not for two reasons. First, overtesting is established through low yield. If physicians

15. See Shanmugam et al. (2015), Obermeyer et al. (2017) for rationale and details.

16. Online Appendix Table A.17 shows the presenting symptom for those ultimately found to have blockage. Nonobvious symptoms (e.g., foot and ankle complaints, nose bleed) are rare but present.

17. We collapse these two tests into one for simplicity (as is reflected in our model). Treating the two tests separately does change our results materially. In Online Appendix A.3, we show the results of performing counterfactuals for each test separately, for example, eliminating all stress tests.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

overtreat, yield will be too high, making it less likely we find overtesting. Second, establishing undertesting does not use information on the yield of testing—only health outcomes—and thus is unaffected.

To flag patients with contraindications $K_{ij} = 1$, we first observe whether they show evidence of poor health prior to visit j (as described above). Second, we observe whether they show evidence of damage to heart muscle by the end of visit j:[18] physicians can note such diagnoses, which is financially incentivized, or we can observe a positive troponin laboratory test suggestive of such problems. If either is present, we assume the physician was aware of possible blockage but decided not to pursue it further because of a contraindication. This assumes all contraindications are measured in our data. In Section IV.C, we explore a broader set of contraindications unobserved in our data but observed by the physician.

Cost-effectiveness is calculated using parameters and assumptions from the literature, summarized in Mahoney et al. (2002) and described in more detail in Online Appendix A.2. Estimates of the benefit of treatment are drawn from clinical trials, which provide estimates of average gains from timely treatment. These trials estimate short-term (e.g., annual) benefits in terms of mortality and morbidity, but total benefits depend on life expectancy. In our model, we abstracted from these considerations for simplicity. Here, to account for actual welfare gains over the life span, we estimate a patient's life expectancy based on their age and individual basket of previsit observed chronic illnesses. We then calculate the life-years a patient would lose from a blockage, both fatal and nonfatal (the latter using a standard discount rate for quality of life losses). Finally, we assume stenting produces a 25% relative reduction in the impact of a blockage; this estimate comes from the most relevant trials, those that randomize testing pathways, for example, immediate versus delayed catheterization. We conduct a sensitivity analysis using a wide range of plausible estimates in Online Appendix A.2. This yields individual-level estimates of the gains from timely treatment, based on the average effect of treatment and the patient's idiosyncratic medical history.

We form an indicator $A_{ij} = 1$ if a patient i experiences a major adverse cardiac event after visit j within a short time window

18. We use this term to denote the medical concepts of infarction and ischemia, a broad category of heart problems including blockage.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

(30 days). The intuition is that blockages have consequences—indeed, this is why we test and treat them—that manifest shortly after onset. We draw on clinical literature that defines these events using the EHR, in a way that shows good agreement with expert judgment after chart review (e.g., Wei et al. 2014). These events fall into three categories: (i) delayed diagnosis and treatment of blockage and diagnosed damage to heart muscle, which we confirm with laboratory biomarkers (positive troponin); (ii) malignant arrhythmia, which we measure using diagnosis codes and cardiopulmonary resuscitation procedures; and (iii) mortality, which we obtain via linkage to Social Security Death Index data. Importantly, apart from mortality, adverse events are only measured if the patient returns to the same health system we study for care. So A_{ij} may be a lower bound on true adverse event rates, relative to widely accepted thresholds from studies that perform active follow-up of enrolled patients. To define objective thresholds for levels of risk that would mandate consideration of testing for blockage, we rely on widely implemented decision rules (e.g., the HEART score of Backus et al. 2010), supported by recommendations from professional societies: 2% over the 30 days after visits. We do not assume such thresholds are optimal; rather, we assume that physicians believe them to be optimal, and thus would not knowingly leave high-risk patients untested. More details are in Online Appendix A.1.C, and additional justification of this threshold based on cost-effectiveness is in Online Appendix A.2.B.

Table I shows that the overall rate of testing is 2.9% of all visits (1.3% with immediate catheterization and 2.0% with stress tests, of which 0.3% subsequently had catheterization, implying a positive stress test). Table II shows that among the tested, the rate of treatment is low: 14.6% (12.9% with stents and 1.8% via open-heart surgery). Among the untested, 27.5% and 11.1% have an ECG and troponin performed, respectively, indicating some suspicion for blockage; 1.2% and 1.9% have explicit evidence of damage to the heart, via the physician's diagnosis ex post and a positive troponin test, respectively. The 30-day adverse event rate is 1.1%.

III.B. Algorithm Design

Our machine learning estimator of risk $\widehat{m}(\cdot)$ is an ensemble model that combines gradient boosted trees and LASSO. It takes

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

TABLE I
SAMPLE SUMMARY STATISTICS

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

Notes. Numbers are fractions unless otherwise noted, reported as mean (std. err.). As a measure of heart disease, past heart disease is the fraction with any diagnosis of heart problems (ischemia), stroke, or peripheral vascular disease prior to the visit. Frequency of individual risk factors (diabetes, hypertension, high cholesterol) is shown, along with the fraction with any of these risk factors.

as its input vector X_{ij} 16,405 characteristics of patient i, observable at the start of visit j.[19] This includes patient demographics; diagnoses, procedures, laboratory results, and quantitative vital signs, measured over the two years prior to the visit; and the symptom recorded at the ED triage desk at the start of the visit. We train estimator $\widehat{m}(X_{ij})$ to predict the yield of testing Y_{ij} among

19. We carefully define X_{ij} to contain only information known to be available to the physician at the time of the decision. We exclude information acquired after triage (i.e., on arrival to the ED): physician notes (which can be completed after the visit) or any data (e.g., ECGs, labs) collected during the visit.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

TABLE II
TESTING OUTCOMES

	All	Tested	Untested
Tested (10 days)	0.029	–	–
	(<0.001)	–	–
Catheterization	0.013	–	–
	(<0.001)	–	–
Stress testing	0.020	–	–
	(<0.001)	–	–
Yield of testing (10 days)	0.004	0.146	–
	(<0.001)	(0.004)	–
Stenting	0.004	0.129	–
	(<0.001)	(0.004)	–
Open-heart surgery	0.001	0.018	–
	(<0.001)	(0.002)	–
Adverse events (30 days)	0.019	0.261	0.011
	(<0.001)	(0.005)	(<0.001)
Diagnosed event	0.016	0.253	0.008
	(<0.001)	(0.005)	(<0.001)
Death	0.004	0.017	0.004
	(<0.001)	(0.002)	(<0.001)
One-year mortality	0.016	0.048	0.015
	(<0.001)	(0.002)	(<0.001)
Physician suspicion (in-ED)			
ECG done	0.294	1.0	0.275
	(<0.001)	(0.004)	(<0.001)
Troponin done	0.131	0.792	0.111
	(<0.001)	(0.005)	(<0.001)
Diagnosed heart damage	0.023	0.391	0.012
	(<0.001)	(0.006)	(<0.001)
Positive troponin	0.025	0.221	0.019
	(<0.001)	(0.005)	(<0.001)
Troponin result (ng/ml)	0.278	0.72	0.124
(if positive)	(0.003)	(0.005)	(0.002)

Notes. Numbers are fractions unless otherwise noted, reported as mean (std. err.). ECG and troponin are low-cost screening tests, done for even a very slight suspicion of blockage. Diagnosed heart damage reflects codes for infarction or ischemia assigned at the end of a visit, and positive troponin indicates damage to heart muscle; both are excluded from calculation of 30-day adverse event rates in untested patients.

the tested, as a close proxy for risk of blockage at the time of an ED visit.[20] To leverage risk information contained in the much larger set of untested patients, we also use predictions on adverse events $A_{ij} = 1$ among untested patients as inputs to the model predicting Y_{ij}. Training happens in a random 75% sample of patients, and

20. To streamline terminology, we refer to this quantity as predicted risk.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

all results below are shown in the remaining 25% hold-out set, except where noted. We split our data set at the patient, not the observation, level, so that all visits from a given patient are assigned to either the training or hold-out set. More details can be found in Online Appendix A.4. Although we cannot share patient-level information to protect privacy, our code repository is publicly available on GitLab.[21]

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

IV. Results

IV.A. Overtesting

Figure I, Panel A shows how well our risk model predicts the yield of testing. In the hold-out set, we sort tested patients into decile bins based on predicted risk. For each bin (x-axis), we calculate the yield of testing (y-axis). Comfortingly, realized yield rises with predicted yield. The algorithm also produces a wide dispersion in realized yields—from 0.01 in the lowest decile to 0.55 in the highest decile.

Figure I, Panel B converts these yields into cost-effectiveness. As in the top panel, patients are sorted by predicted risk, but this time into quintile bins (x-axis).[22] The y-axis now shows the implied cost-effectiveness of testing patients in a bin, in units of thousands of dollars per life-year. The y-axis shows a commonly used threshold for judging cost-effectiveness, \$150,000, as well as the cost-effectiveness of selected other procedures for comparison. This illustrates a great deal of inefficient testing. The bottom bin of tests is extremely cost-ineffective: \$1,352,466 per life-year. For comparison, biologics for rare diseases (some of the least

21. https://www.gitlab.com/labsysmed/zolab-projects/stressed_ensemble

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

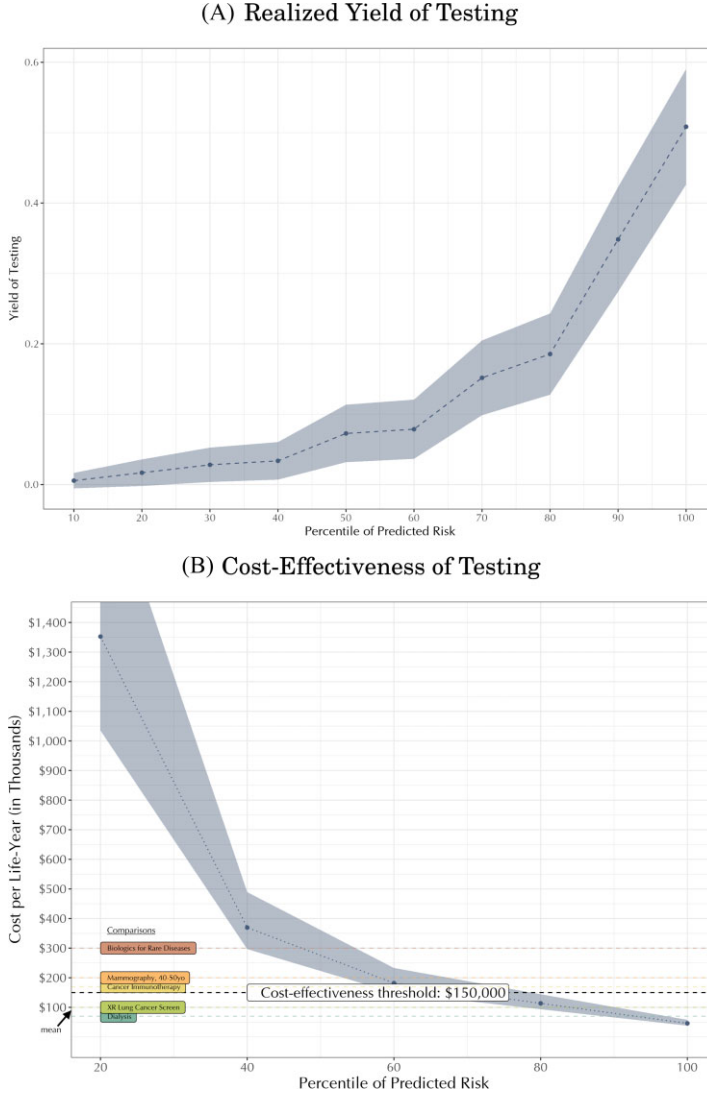

FIGURE I

Yield and Cost-Effectiveness of Testing in Tested Patients

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

cost-effective technologies that health systems sometimes pay for) are typically estimated at around $300,000 per quality adjusted life-year.[23] Even the second-lowest bin is very cost-ineffective at $318,603 dollars per life-year.

With these data, we can calculate a precise policy counterfactual as described in Lemma 1: dropping individual tests whose cost-effectiveness predictably falls below a threshold. For example, at a $150,000 life-year valuation, we would drop 62.4% of the lowest-value tests, with a combined cost-effectiveness of $265,114 per life-year.[24] These results only deal with one kind of counterfactual: eliminating the particular tests physicians decided to do (i.e., stress tests or catheterizations) on patients in a given risk bin. Since we have two types of tests, Online Appendix A.3 also explores other counterfactuals. A notable finding is that stress testing (as opposed to catheterization) is so low-value that eliminating it altogether would improve welfare, as has been previously suggested (Prasad, Cheung, and Cifu 2012). Taken together, the results in Figure I and these policy counterfactuals suggest a great deal of overtesting.

IV.B. Undertesting

At the same time, testing in the high-risk bins appears very cost effective. Table III, column (2) shows that in the highest-risk quintile bins, tests cost only $46,017 per life-year, comparable with cost-effective interventions like dialysis. In column (3), we show testing rate by predicted risk for all patients (for comparability, these bins are formed using the same bin cutoffs used in the tested set, so they are not equally sized). We see that physicians do test higher-risk patients more. But strikingly, many high-risk patients go untested—only 38.3% in the top bin are actually tested.

Of course, this only tells us that the physician and the model disagree, not who is right.[25] The physician has access to a host of

23. Online Appendix A.2 shows that these estimates are not sensitive to the particular choice of parameters in our analysis, and in particular hold over wide ranges of possible treatment effect sizes.

24. In Lemma 1, establishing that a set of patients were overtested also required that this set had lower than average yield. That condition also holds here, as is seen in Table III, where the bottom six deciles have yield well below the average of 0.146.

25. To some extent, any two models of risk—even very good ones—may differ due to noise. So perhaps any discrepancies we see between the physician and the model could simply be the consequence of comparing two well-fit models to each

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

TABLE III
REALIZED YIELD, COST-EFFECTIVENESS, AND TESTING RATE

	Yield rate (std. err.) (1)	Cost-effectiveness ($) (lower–upper bound) (2)	Test rate (std. err.) (3)
Full sample	0.146	89,714	0.029
	(0.004)	(74,152–113,543)	(<0.001)
By risk bin			
1	0.011	1,352,466	0.012
	(0.006)	(1,034,814–1,951,515)	(<0.001)
2	0.036	318,603	0.017
	(0.01)	(257,296–418,265)	(0.001)
3	0.07	192,482	0.047
	(0.014)	(157,552–247,314)	(0.002)
4	0.168	114,146	0.088
	(0.02)	(94,154–144,914)	(0.004)
5	0.429	46,017	0.383
	(0.026)	(38,178–57,907)	(0.016)
N	1,784	1,784	61,965

Notes. Yield of testing (1) and cost-effectiveness of testing (2) in the tested, and test rate across all visits (3), by quintile bins of predicted risk. Risk bin cutoffs are defined in the tested population, so bins here are equally sized in columns (1) and (2), but not in (3) (which describes the entire population—tested and untested). Lower–upper bounds on cost-effectiveness are defined by a range of plausible estimates of the effect of testing on health, from randomized trials.

information unavailable to the model: how the patient looks, what they say, or crucial data in the ED such as X-rays or ECGs. These data elements are likely to be predictive of yield; if they are also predictive of testing, this private information will create selection bias: untested patients will have far lower yield than predicted based on observables.

Because we lack test results on the untested, we have no way to quantify the magnitude of the problem. But a simple calculation suggests a large bias. The hold-out set has 266 positive tests. Taking model predictions at face value would imply 10 times as many positives (2,738) were we to test the predicted high-risk untested—implausibly large. To show the role of private information more directly, Online Appendix A.5 incorporates data from ECGs, observed by the physician but not routinely

other. In Online Appendix Figure A.11, we compare two machine learning models fit on separate samples of our training set and find that these correlate much more strongly than the model and the physician do. More important, we perform a variety of tests that directly test for error, both in the sense of welfare-enhancing counterfactuals and specific behavioral errors.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

observable in health data sets, into risk predictions.[26] For patients with ECG data available, we show that several ECG features (e.g., ST-elevation, ST-depression) predict both the physician's test decision and the yield of testing, conditional on $\hat{m}(X)$: physicians are using these data effectively. We then directly incorporate the ECG waveform into new risk predictions, via a deep learning model. This decreases model-predicted risk for 97.5% of patients, and 100% of the highest-risk untested. So the model without the ECG was significantly overestimating the risk of the untested patients. Of course, the ECG is just one of many critical variables we do not (and cannot) observe.

Following Lemma 1, we look for evidence of undertesting in the form of adverse events resulting from untreated blockages, in the 30 days after visits. Among all eligible untested patients, the rate of adverse events is 1.1%, well below the 2% clinical threshold, implying (reassuringly) that testing all untested patients does not make sense.[27] Figure II shows these adverse-event rates (y-axis) by decile bins of predicted risk. Again, for comparability we use bin cutoffs defined in the tested, meaning bins are of unequal sizes in the untested: in particular, because the untested are lower risk than the tested, bin size decreases in risk. Panel A shows that patients in high-risk bins have very high 30-day adverse-event rates. For example, the highest-risk bin contains 0.15% of the untested, 15.6% of which go on to have an adverse event. The second-highest-risk bin contains 0.75% of the untested and has an adverse event rate of 6.81%; together the top two bins have an adverse event rate of 8.26%. In fact, the crossover point where the adverse event rate becomes statistically indistinguishable from the 2% threshold is the sixth risk bin, which means that the top four bins—which make up 6.9% of the untested—all have high enough adverse event rates to merit consideration for testing under current guidelines.

These adverse events are not simply billing codes, which might exaggerate the incidence of actual health problems, due

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

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

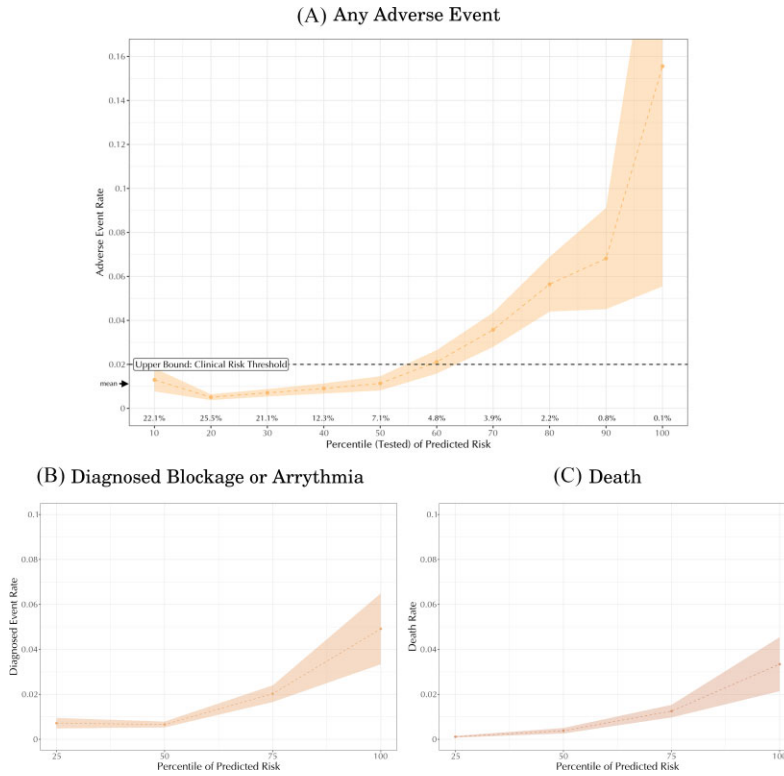

FIGURE II

Adverse Events in Untested Patients (30 Days after Visits)

Thirty-day adverse event rates among untested patients (y-axis), by decile bins of predicted risk (x-axis). Risk bin cutoffs are defined in the tested population, so bins here are not equally sized: the percent in each bin is shown above the x-axis. Panel A shows the total adverse-event rate (the top of the highest 95% confidence interval is truncated). The horizontal line shows the 2% threshold above which testing is recommended by clinical guidelines; the highest-risk 14% (top six bins) have a rate significantly above 2%. The bottom panels show two subset categories of adverse events that make up the total: (B) diagnosed adverse events (heart damage, confirmed with laboratory biomarkers; and cardiac arrest), (C) death (via linkage to Social Security data); bins here are quartiles of predicted risk (because outcomes are less frequent).

to incentives to overtest or treat. Panel B shows the subset of adverse events related to diagnosed blockage, all confirmed with biomarker evidence of damage to the heart muscle (positive troponin laboratory results), as well as dangerous arrythmias (ventricular fibrillation and tachycardia, or procedure codes for

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

defibrillation or CPR). In the highest-risk bin, 4.9% have one of these events. Panel C shows 30-day mortality. The highest-risk bin experiences death at a rate of 3.3%, comprising nearly half (45%) of all adverse events in this bin. These data alone suggest a great deal of undertesting. However, there is a potential confound, which we address next.

IV.C. Accounting for Differences in Treatment Benefits

These high adverse event rates establish that predicted high-risk patients who go untested are indeed high risk. But it does not establish that failing to test them was a mistake. Adverse events rule out private information by physicians about risk, but not private information about the suitability of treatment. It is possible that physicians recognized these patients as being high risk, but also recognized them as having lower return to treatment and chose not to test them for that reason. In particular, we may have mismeasured K_{ij}. In excluding patients $K_{ij} = 0$ from our sample (by excluding those with prior ill health and by excluding untested patients in whom the physician appears to suspect heart problems), our measure K_{ij} may have failed to capture other elements of K that the physician observes. One fact provides prima facie evidence that these unobservables are not large: the average age of the untested we flag for testing is 58.5 (close to the mean age of the tested, 57.8), whereas the average age of those with observed contraindications is 68.5. At least on this crucial observable, the high-risk untested look more like the tested than the too frail to test.

To address this problem more thoroughly, we use a clinical fact. When physicians suspect a blockage, even if the patient is ineligible for testing or treatment, there are still important actions they can and must take. At a minimum, everyone the physician suspects of a blockage will be given an ECG—a low-cost, noninvasive test. Even for treatment-ineligible patients, the ECG guides medications (e.g., blood thinners) and decisions about intensity of monitoring (e.g., whether to admit to the ICU). Similarly, the troponin blood test will also be checked, as it provides critical information on the nature and extent of any blockage. So if we remove patients with an ECG or troponin from our calculations, we will have removed all patients in whom physicians had even the slightest suspicion of a heart

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

problem, leaving us with a pool of unsuspected patients.[28] Within the remaining unsuspected pool, we recalculate the adverse event rate. If the high adverse event rates in the whole population are due to physicians knowingly leaving some high-risk patients untested, because they are unsuitable for treatment, then this unsuspected pool should have a very low adverse event rate, and specifically the rates should be below the clinical threshold for testing.

The top two panels of Figure III first show the fraction of all patients who are both untested, and did not receive an ECG (Panel A) or troponin (Panel B), by quartile bin of predicted risk. As expected, higher-risk patients are on average perceived as such by physicians: they are less likely to be untested and lack one of these tests. Though decreasing, the fractions nonetheless remain substantial in the highest-risk bin: 19.1% are untested and lack an ECG (versus 77.7% in the lowest-risk bin), and 41.2% are untested and lack a troponin result (versus 93.3% in the lowest-risk bin). The bottom two panels show the adverse event rates in only these untested patients without an ECG or without troponin. For the highest-risk untested patients without such suspicion for heart attack, adverse-event rates remain high: 4.3% in those without an ECG, and 6.6% in those without a troponin. These rates are 3.2 percentage points (std. err. 1.3) and 1.2 percentage points (std. err. 1.1) lower than the 7.5% rate in the full population above, respectively; but they still significantly exceed the clinical threshold for testing of 2%.[29] Together, these results suggest that physicians do have private information both about the risk of blockage and about suitability for treatment—but that even after accounting for them, there is still substantial undertesting.

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

29. Online Appendix A.6.C describes another sensitivity analysis, in which we eliminate patients who were admitted to the hospital with an uncertain diagnosis (e.g., those with a symptom-based diagnosis code like chest pain, as opposed to a specific disease), in whom physicians may have latent concern for blockage. When we calculate adverse event rates in the remaining patients—those in whom the physician felt sure enough to assign an alternative diagnosis other than blockage, and those discharged home from the ED and thus at very low risk of serious problems—we find similar results: a rate of adverse events equal to 8.43% in the highest-risk bin, as opposed to 8.26% in the full population.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

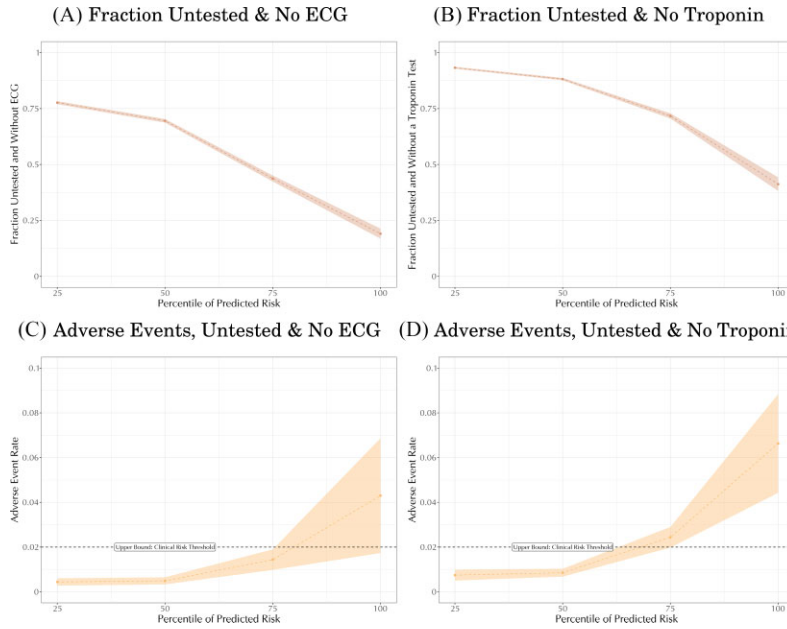

FIGURE III

Adverse Events in Untested and Unsuspected Patients

Top panels: fraction of patients in whom physicians do not appear to suspect blockage. Panel A shows the fraction untested and lacking an electrocardiogram (ECG); Panel B shows the fraction untested and lacking a troponin laboratory test. Both ECG and troponin are low-cost tests used to screen for blockage; they are done even in patients who may be ineligible for invasive treatment. Fractions are shown by quartile risk bins, with bin cutoffs defined in the tested population (so bins here are not equally sized). Bottom panels: rate of 30-day adverse events (diagnosed events and death) after visits (y-axis), by bin of predicted risk (x-axis), among untested patients lacking (C) an ECG, and (D) a troponin. The horizontal line shows the clinical threshold above which testing is recommended by clinical guidelines.

IV.D. Natural Experiment

Although these data provide clear evidence of undertesting, this evidence is indirect, based on clinical thresholds. It would be reassuring to have more direct evidence that testing these untested high-risk patients would affect their health. Ideally, we would measure the effect of testing some high-risk patients at random, and see if in fact mortality and long-term adverse event rates decrease significantly. While such an experiment is beyond

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

the scope of this article, we can exploit natural variation in our data that might serve as a (limited) proxy for it.[30]

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

Our data do not track the exact identity of the triage team, but we know the times at which shifts begin and end. This lets us calculate the average testing rate of all other patients seen on a shift, $\bar{T}_{-j}$, to instrument for whether patient-visit j is tested. For this to be a valid instrument, we assume that (i) the triage shift affects long-term health outcomes only through testing, and (ii) patients are balanced on unobservables across shifts; we discuss both assumptions below. We perform this analysis on a slightly different sample than used so far. To maximize power, we use the full data set, not just the hold-out. To avoid overfitting, we use fivefold cross-validation to predict risk. In addition, to address nonindependence of health outcomes across visits, we restrict the sample to each patient's first visit.[31]

Overall, there is reasonable variation in likelihood of testing across shifts: for example, a patient in the highest-risk bin arriving on a Monday evening is 18% more likely to be tested by the highest- (19.9%) versus lowest-decile (16.8%) shifts. Regressing a visit's test (T_j) on the leave-one-out shift testing rate

30. In the context of the framework, the natural experiment measures the health effect of testing due to the treatments that result from that testing. As such, it measures the joint effect of the increased propensity to test and the treatment effect conditional on a positive test. This will tell us whether the resulting health benefits are above or below what would merit testing.

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

($\bar{T}_{-j}$), controlling for time fixed effects (year, week of year, day of week, and hour of day) and patient risk, we find that a one standard deviation increase in shift testing rate (2.3 percentage points) increases individual testing probability by 0.19 percentage points (std. err. 0.06), or 6.7% of the base test rate (see Online Appendix Table A.12).[32]

Figure IV shows how patient observables compare across shifts. The top panel shows the results of regressing a pretriage variable X_j on the shift testing rate. We do find statistically significant differences in predicted risk across triage testing rates ($p = .051$), but they are very small in magnitude: a 1 standard deviation increase in $\bar{T}_{-j}$ implies a 0.007 standard deviation difference in predicted risk. But reassuringly, we find no statistically significant difference when we test for differences in predicted risk nonlinearly (by risk bin), nor in age, sex, self-reported race, income, or risk factors for heart disease. Together, these results suggest that observables are (largely) balanced across shifts. In the bottom panel, we plot for each shift the average testing rate for all patients who arrive in that shift (in percentile terms, x-axis) and the average predicted risk of those patients (y-axis). We see that at every level of testing rate, there is large variability in predicted risk.

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

We then measure the overall effect of testing on health, as measured by long-term adverse events A_j^ℓ, by estimating:[33]

$$(1) \qquad A_j^\ell = \beta_0 + \bar{T}_{-j}\beta_1 + \widehat{m}(X_{ij})\beta_2 + \text{Time Controls}_j\beta_3 + \epsilon_j.$$

32. In Online Appendix Table A.11, we also rule out that hospital capacity constraints on testing facilities might be reducing the likelihood of testing, by showing that a visit's likelihood of testing is not affected by the number of tests done in the 12–28 hours before the visit.

33. We measure some outcomes over the 31–365 days after ED visits because tested patients are mechanically more likely to be diagnosed with heart problems than untested patients, simply by virtue of being in the hospital for testing. By contrast, our mortality data come from linkage to Social Security data, and so do not suffer from this difference in ascertainment.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

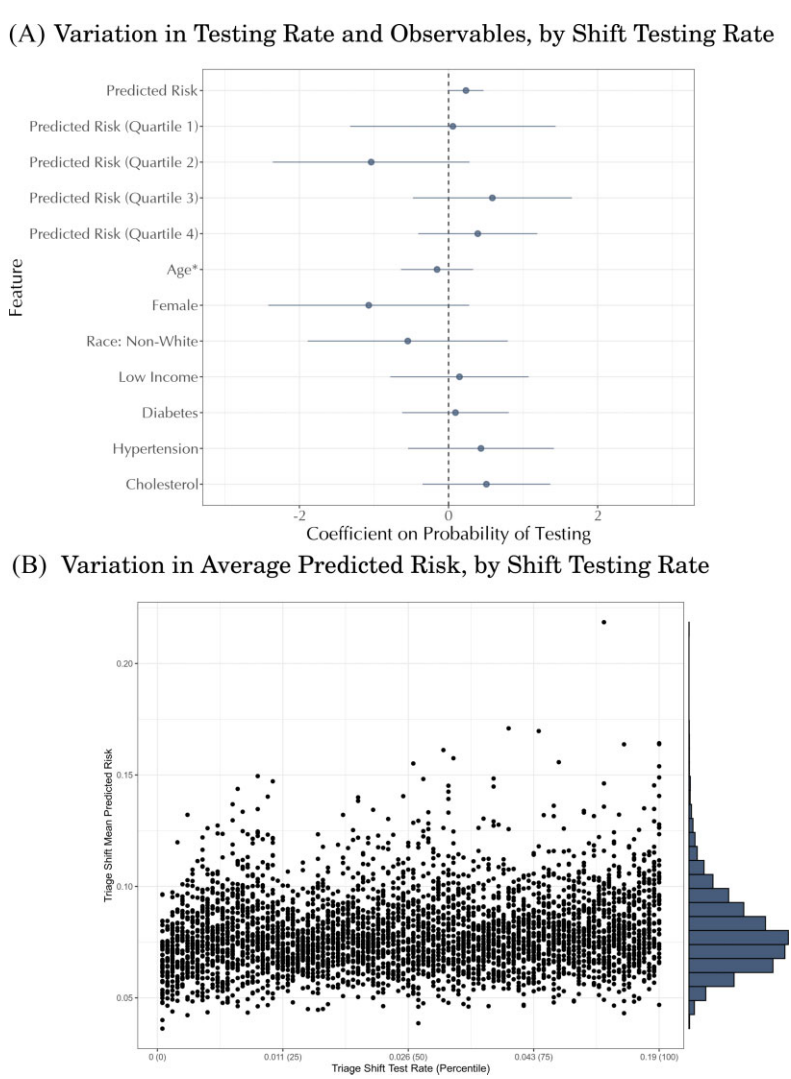

FIGURE IV

Balance and Risk Variation across Triage Shifts

Panel A shows balance checks in a quasi-experiment, in which patients arriving during different triage shifts are tested at higher or lower rates. Each point shows the coefficient and confidence interval on the leave-one-out shift testing rate ($\bar{T}_{-j}$), from a regression of a given pretriage variable on $\bar{T}_{-j}$. Panel B plots, for each shift, the average testing rate for all patients who arrive in that shift (in percentile terms, x-axis) and the average predicted risk of those patients (y-axis). Each point represents one of 3,951 shifts in our data set, and the density plot on the right shows the overall distribution of mean risk. *Age is divided by 100 for scale.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

TABLE IV

EFFECT OF TESTING ON HEALTH, USING SHIFT TESTING VARIATION

	Diagnosed event (31–365) (1)	Death (31–365) (2)	Death (0–365) (3)
Panel A: Average effect			
Predicted risk	0.05***	0.15***	0.25***
	(0.005)	(0.01)	(0.01)
Shift test rate	0.02	0.005	0.005
	(0.01)	(0.01)	(0.02)
Observations	123,289	123,289	123,289
Panel B: Heterogeneous effect by risk			
Predicted risk	0.06***	0.17***	0.27***
	(0.01)	(0.01)	(0.01)
Shift test rate	0.04**	0.04**	0.04*
	(0.02)	(0.02)	(0.02)
Predicted risk	− 0.25*	− 0.49***	− 0.43**
× Shift test rate	(0.15)	(0.17)	(0.20)
Observations	123,289	123,289	123,289
Outcome rate	0.018	0.012	0.016
Outcome rate, top risk bin	0.027	0.046	0.077

Notes. Panel A: Regression of diagnosed adverse events (column (1)) and death over days 31–365 after visits (column (2)) on leave-one-out shift testing rate. We use 31–365 days because tested patients are mechanically more likely to be diagnosed with heart problems than untested patients in the first 30 days. Our mortality data, by contrast, do not suffer from this difference in ascertainment, so death over the full year after visits is also shown (column (3)). Panel B: The same regression, but with an additional interaction term that allows the effect of testing to vary by predicted risk. Outcome rates, overall and in the top risk quintile, are shown below. Controls for time (fixed effects: year, week of year, day of week, and hour of day) and patient risk are included but not shown. This sample includes only patient i's first visit j, to address nonindependence of outcomes across visits, so the sample size is reduced. * $p < .1$; ** $p < .05$; *** $p < .01$.

That is, we regress adverse events in the year after visits on shift testing rates, controlling for time fixed effects (year, week of year, day of week, and hour of day) and patient risk. Table IV, Panel A shows that on average, we find no statistically significant effects on health outcomes. Neither diagnosed adverse events from day 31–365 after visits (column (1)) nor death, whether measured over the same period as diagnosed events (column (2)) or over the full year after visits (column (3)), are affected.

As before, the average effect may conceal a great deal of heterogeneity: undertesting is not universal but only in high-risk patients. So we reestimate equation (1), but include an interaction term $\bar{T}_{-j} \times \widehat{m}(X_j)$, which allows the effect of testing to vary by predicted risk. Table IV, Panel B shows this interaction term to be large, negative, and significant, indicating lower rates of

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

diagnosed events and death in higher-risk patients. To scale this coefficient, the implied reduction in one-year mortality for the highest-risk quintile is 2.6 percentage points (34%) if they arrive on the highest- versus lowest-testing shifts. This confirms that physician private information about treatment heterogeneity cannot account for our findings: increased testing improves health in high-risk patients. It also provides some reassurance regarding the exclusion restriction in our experiment: if triage affected long-term outcomes in ways unrelated to testing for blockage, we would expect to see broader effects, not just among the predicted high-risk for blockage. We emphasize that this does not imply that all high-risk untested patients would benefit from testing: we are constrained by the extent of variation in testing rates in our quasi-experiment and can say nothing about patients who are never tested (i.e., even in the highest-testing shifts).

We use these estimates to simulate counterfactuals that bound the extent of undertesting. We first estimate a random effects model of a shift's testing rates and group shifts into quartiles based on its random effect.[34] Suppose we know a predicted risk bin has positive benefits from testing. Our counterfactual assumes all such patients are assigned to the highest-testing shifts: so the difference between a patient's actual shift testing rate and highest-quartile shift test rate is counted as undertesting. The key assumption is which risk bins have positive benefits from testing.

We take two approaches. First and most conservative, we assume only those with significant one-year mortality reductions from testing qualify, which based on Table IV includes only the top risk quintile. (Recall that bins are defined using bins in the tested, so the top quintile is far less than 20% of the whole population.) Reassigning all patients in this bin from their actual shift (mean 18.1% test rate) to the highest-quartile shift (32.3% test rate), generates additional tests equal to 0.48% of all untested patients or 15.6% of the tested set. A second approach allows for other testing benefits beyond decreasing one-year mortality, for example, reductions in immediate heart attack (size and extent), as well as longer-term outcomes. To simulate this, for each risk

34. The leave-one-out shift testing rate, while useful for identification of the effect of testing, does not capture the full variation in observed testing rate across shifts. Online Appendix A.7.C contains more details on the model, which controls for the same vector of time variables and patients' predicted risk as above.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

bin, we take the cost-effectiveness estimates from the tested and (naively) apply them to the untested. By testing patients who appear to be cost-effective based on risk, we would add new tests equal to 3.0% of the set of all untested, and 99.5% of the current tested set.[35]

Taken together, the evidence tells us three facts about high-risk untested patients, all suggesting they ought to have been tested. First, they go on to have high adverse-event rates of the kind that suggest undiagnosed blockage. Second, physicians do not appear to have recognized their risk: many were not screened with simple tests given to everyone suspected of any heart problems (ECG or troponin), but nonetheless had high adverse event rates. Finally, plausibly exogenous increases in testing improve their health, but not the health of lower-risk patients. Each finding has its limitations, but together, they make the case that testing high-risk untested patients would increase welfare as strongly as possible without a randomized trial.

IV.E. Nationally Representative Data

These results come from a single hospital. To check their generality, we replicate them in a nationally representative 20% sample of Medicare fee-for-service patients, from January 2009 through June 2013. These data are limited in several important ways. Because they are based on insurance claims, not EHR data, they contain very limited patient information. For example, we do not have ECGs, lab values, or other biomarkers, nor do we have arrival time and shift timing data that would let us re-create our natural experiment. These caveats aside, these data do let us replicate our estimates of over- and undertesting from Sections IV.A and IV.B. Applying similar exclusions to those used in the single-hospital data, we arrive at a final sample of 4,425,247 Medicare visits by 1,602,501 patients, of whom 4.4% were tested. Of the tested, 12.4% received treatments. Of the untested, 5.3% had 30-day adverse events. This higher rate reflects the older and sicker Medicare population, but also our inability to confirm diagnosis codes with biomarker evidence of heart attack as above.

35. Note that irrespective of the risk threshold we choose, this strategy still respects the large amount of physician private information we document: we do not propose that 100% of patients in a high-benefit risk bin should be tested. The never-tested—even those in high-risk bins—may have unobservables that lead them to be lower risk. Our strategy simply shifts the testing rate from the current rate to the maximum rate we observe for a given risk bin.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

Online Appendix Figure A.7 shows that yield of testing and cost-effectiveness both increase in predicted risk (as in Figure I), with many tests being predictably cost-ineffective. We also find many high-risk untested patients with adverse-event rates above clinical thresholds. Online Appendix Figure A.8 shows that 3.8% of the highest-risk patients are diagnosed with an adverse event, and an additional 1.5% die (as in Figure II). In summary, we find both overtesting (52.6% of all tests) and undertesting (at least 17.9% of the tested).[36]

V. Why Do Physicians Make Testing Errors?

We have shown that physicians mispredict: they test predictably low-risk patients and fail to test predictably high-risk patients. In this section, we try to better understand the nature of physician misprediction. To do so, we examine how physician testing decisions deviate from predicted risk. Our approach builds on a long tradition of research comparing clinical judgment to statistical models as a way to gain insights into decision-making, often among physicians (Dawes, Faust, and Meehl 1989; Elstein 1999; Redelmeier et al. 2001; Ægisdóttir et al. 2006), as well as the clinical literature on diagnostic error (Croskerry 2002; Graber, Franklin, and Gordon 2005; IOM 2015). We view this as exploratory: a way to shed light on potential psychology at work, rather than to structurally estimate a specific model of physician decisions.

V.A. Boundedness in Physician Judgments

One reason physicians may make errors is that the optimal risk model is quite complex: our own machine learning model uses 16,405 variables. Bounded rationality may lead them to use a

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

simpler approximation. Such simplification is analogous to regularization in machine learning (Camerer 2019). To avoid overfitting, algorithms do not pick the model that fits best in sample. Instead they estimate a best-fit model for each level of complexity, then choose a complexity level by asking which of these best-fit models produces best out-of-sample fit. To study physicians, we use this same set of best-fit models for each complexity. But we ask which model complexity best predicts physician choices, not out-of-sample risk. If physicians are boundedly rational, the model that best predicts their choices should be simpler than the one that best predicts actual risk, measured by yield of testing.

We implement this procedure using the LASSO model of risk, one component of our full ensemble model, because it has a straightforward measure of complexity: the number of nonzero coefficients included in its linear model.[37] For $k \in [0, 1{,}500]$ we train and retain the set of best-fit LASSO models that has exactly k nonzero coefficients.[38] In our hold-out set, we correlate each model with test outcomes and testing decisions. Two caveats are worth noting. First, we do not assume anything about the model selection properties of LASSO: the particular variables the LASSO chooses is somewhat arbitrary in the setting of correlated, noisy input variables. We are interested only in the complexity of these models, which is likely a more stable quantity (Mullainathan and Spiess 2017). Second, we can only focus on the variables in our data: so we only test hypotheses related to boundedness on observables, not on the variables physicians may use that are unobservable to us.

Figure V visually displays the results of this exercise. On the x-axis is k, the measure of complexity. On the y-axis is R^2, a measure of goodness of fit (though our results are not specific to this setup: Online Appendix Figure A.12 shows similar results with AUC instead of R^2, trees instead of LASSO, and the Medicare population). The gray line shows, at each level of complexity, how well a model predicts out-of-sample risk: R^2 increases at first, then decreases as additional variables lead to overfitting. The yellow line shows how well the same model predicts physician testing

37. Though this is a suitable ex post measure, ex ante this is produced by using L_1 regularization.

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

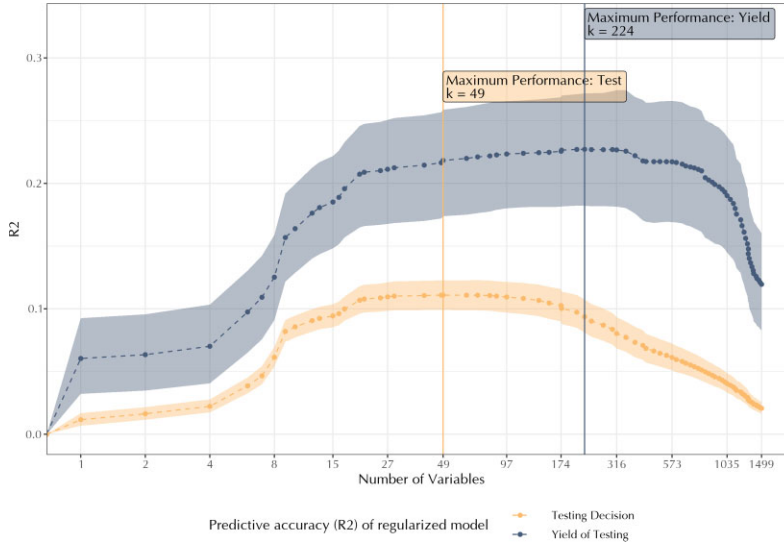

FIGURE V

Explanatory Power of Simple versus Complex Models of Risk

Using a LASSO model of predicted risk (part of our full ensemble risk model), we preserve all risk models along the regularization path for $k \in [0, 1500]$: the best-fit linear model that uses at most k nonzero coefficients. The x-axis shows k, the number of variables retained as the regularization penalty is decreased, moving from left to right. The y-axis shows the explanatory power of these risk models of varying complexity for physician testing decisions (dark gray line), and patient risk (yield of testing: yellow line), measured by R^2. The 95% confidence intervals are the shaded intervals, calculated by bootstrapping. The two vertical lines show the complexity of the model that explains the most variance in physician decisions (left, at $k = 49$) and risk (right, at $k = 224$).

decisions. Here we see in part a similar pattern: R^2 increases with complexity, then decreases. Importantly, the two curves hit their peaks at very different levels. For physicians, the empirical optimum is at 49 variables, and for risk it is at 224 variables. The model that best predicts actual risk is much more complex than the one that best predicts test decisions.

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

TABLE V
EVIDENCE FOR PHYSICIAN BOUNDEDNESS

	Test		Yield	
	(1)	(2)	(3)	(4)
Predicted risk, simple ($k = 49$)	1.357*** (0.015)	1.358*** (0.016)	1.528*** (0.068)	1.319*** (0.081)
Incremental risk, complex ($k = 224$)		−0.005 (0.033)		1.099*** (0.236)
Observations	61,821	61,821	1,834	1,834
R^2	0.111	0.111	0.218	0.227

Notes. Tests of the explanatory power of two versions of predicted risk, for physician testing decisions and patient risk (yield of testing). We first identify the simple risk model of complexity that explains the most variance in physician decisions (with $k = 49$, here labeled *Predicted risk, simple*). We then subtract this prediction from the risk model of complexity that explains the most variance in patient risk (with $k = 224$, here labeled *Incremental risk, complex*). Columns (1) and (3) show how the simple risk model predicts both test and yield alone. Columns (2) and (4) then add the complex model's incremental contribution to predicted risk. * $p < .1$, ** $p < .05$, *** $p < .01$.

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

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

If physicians rely only on a simple model of risk, we expect two things to be true. First, $\beta_2 = 0$: complex risk should not predict testing decisions. Second, $\gamma_2 > 0$: complex risk should predict yield. Table V, columns (1) and (3), show how the simple risk model alone predicts both test and yield; columns (2) and (4) show the addition of complex risk. In column (2), as expected, complex risk is not predictive of testing conditional on simple risk—the coefficient is both very small and statistically insignificant. In column (4), by contrast, complex risk is predictive of yield and highly significant. So physicians do in fact appear to rely on too simple a model of risk.[39]

These results provide suggestive evidence that physicians are boundedly attentive: they only pay attention to some variables. But how accurately do they weigh the variables they attend to? Figure VI shows, for the 49 variables in $\widehat{m}_{\text{simple}}(X_{ij})$, their

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

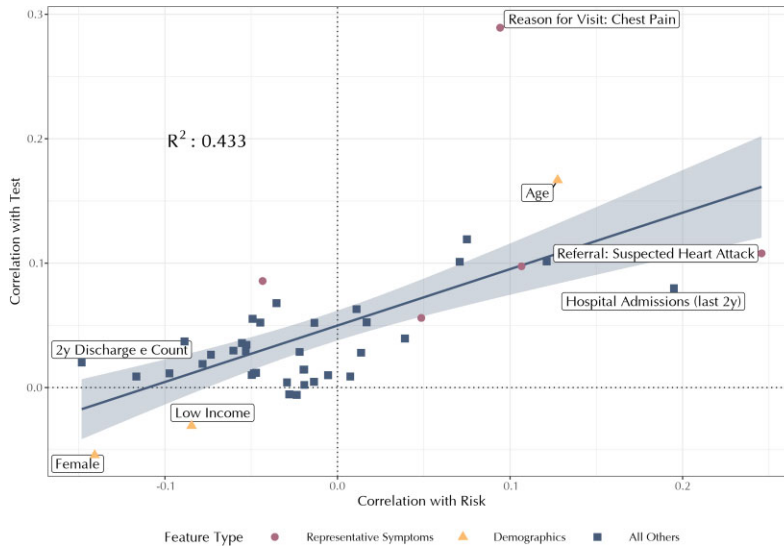

FIGURE VI

Simple Risk Variables: Correlation with Testing and Predicted Risk

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

correlation with test outcome (x-axis) and test decision (y-axis).[40] We see a tight, strongly positive relationship ($R^2 = 0.433$). While far from proof of rationality, this does suggest that physicians (mostly) correctly weight the variables they do use.

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

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

quintile). Likewise, among the undertested, 74.2% come from the bottom quintile, meaning their simple risk is much smaller than their actual risk (compared with 7.4% in the top quintile). Boundedness thus appears to be quantitatively important as well for misprediction. Physicians identify a handful of good risk predictors that they use, if not perfectly, at least modestly well; at the same time, they neglect many other variables which, while individually small, together provide much explanatory power.

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

V.B. Biases in Physician Judgments

Figure VI, while largely consistent with bounded rationality, also hints at another phenomenon: physicians might over- or underweight specific variables. In particular, a suggestive example is "Reason for visit: chest pain," a clear outlier. A complaint of chest pain does correlate with risk, but it correlates even more with testing. This indicates that those with chest pain may be tested at rates above and beyond what is justified by their (heightened) risk.[41]

Chest pain has two features that make it particularly interesting from a behavioral point of view, suggesting two broader behavioral hypotheses for why an input might be overweighted. First, it is highly salient (Tversky and Kahneman 1974; Bordalo, Gennaioli, and Shleifer 2012). Second, it is highly representative of blockage: it is a (perhaps the) stereotypical symptom in textbooks and in public understanding (Bordalo et al.

41. Conditional on predicted risk, patients with chest pain are 16 percentage points (578%) more likely to be tested. Online Appendix Table A.15 shows that for the 10 most common symptoms, 9 significantly predict testing after conditioning on predicted risk, including chest pain and shortness of breath (large and positive), and several other smaller negative predictors (e.g., abdominal pain).

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

2016). This motivates our exploration of bias: we ask whether variables that are either salient or representative are generally overweighted.

We study these hypotheses in turn, using a similar empirical approach. To assess whether physicians are biased in their use of some subset of variables $\mathcal{W}$, we create a new risk predictor that uses only those variables in $\mathcal{W}$. Except for the restriction on input variables, this estimator, $\widehat{m}_{\mathcal{W}}$, is built in the training set exactly the same as the original risk predictor, and for simplicity of notation takes the same input X_{ij} but ignores the variables not in $\mathcal{W}$. In the hold-out set, we first regress yield on full risk (our usual risk predictor $\widehat{m}(X_{ij})$) as well as this limited risk model $\widehat{m}_{\mathcal{W}}(X_{ij})$, analogous to equation (2).[42] We do this to verify that, as expected, conditional on full risk, $\widehat{m}_{\mathcal{W}}$ does not provide additional information. Then, as our test of whether $\mathcal{W}$ is misused, we regress the test decision T_{ij} on full risk $\widehat{m}(X_{ij})$ and $\widehat{m}_{\mathcal{W}}(X_{ij})$. If physicians overweight the variables in $\mathcal{W}$, the coefficient on $\widehat{m}_{\mathcal{W}}(X_{ij})$ should be positive; if they underweight, it should be negative.[43]

1. *Symptom Salience.* Building on the chest pain insight above, we implement this procedure first for symptoms: the most salient and immediate thing the physician sees about a patient, often stressed in medical education and vignettes. Table VI, column (1) shows the results of regressing testing on the full risk predictor; column (2) then adds the new symptom-only risk predictor.[44] We see here that the risk from symptoms is additionally predictive of testing, suggesting that symptoms as a category are overweighted.[45]

42. All regressions control for a vector of risk bins, as well as linear risk, to account for nonlinearity of risk in predicted risk. We show the linear coefficient but omit the others for simplicity.

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

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

45. Abaluck et al. (2016), although they lacked data on symptoms at the visit itself, found that patients with past symptom-based diagnoses were overtested, consistent with a similar bias.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

TABLE VI

SYMPTOM SALIENCE AND REPRESENTATIVENESS

	Test				
	(1)	(2)	(3)	(4)	(5)
Predicted risk, full	0.872***	0.715***	0.756***	0.619***	0.755***
	(0.053)	(0.049)	(0.061)	(0.045)	(0.066)
Predicted risk, subsets					
All symptoms		0.888***	0.860***	0.273***	
		(0.052)	(0.057)	(0.061)	
Representative symptoms				1.283***	
				(0.121)	
Demographics			0.139***		
			(0.031)		
Prior diagnoses			0.046**		
			(0.021)		
Prior procedures			−0.053*		
			(0.030)		
Prior lab results and vital signs			−0.209***		
			(0.019)		
Physician experience					
Experience (years)					−0.0005**
					(<0.001)
Experience × risk					0.011***
					(0.005)
Observations	61,938	61,938	61,938	61,938	55,777
R^2	0.084	0.106	0.113	0.118	0.082

Notes. Column (1) regresses testing on our usual predicted risk measure $\hat{m}(X_{ij})$. Column (2) adds a risk predictor formed using only symptom inputs. Column (3) adds risk predictors to column (2), formed using other input categories. Column (4) adds another risk predictor to column (2), formed from only nine representative symptoms. Column (5) regresses testing on predicted risk and physician experience (linear and interacted with risk). All models also control for nonlinear risk terms (not shown). Similar regressions with yield of testing as the dependent variable are shown in Online Appendix Table A.18, confirming that none of these variables are predictive over and above $\hat{m}(X_{ij})$. * $p < .1$, ** $p < .05$, *** $p < .01$.

We expand this exercise to the entire universe of inputs. We form a set of risk predictors, one for each subset of variables, grouped into the following categories: demographics, prior diagnoses, past procedures done on the patient, and prior labs and vital signs. The categories are formed to reflect coherent types of inputs physicians may treat differently. For example, medical case reports and pedagogy use a standard structure, stressing age, sex, and symptoms (e.g., "A 43-year-old man with chest pain," as in the *NEJM*'s Case Records). So we conjectured that demographics and symptoms would be highly salient and thus overweighted. By contrast, the complex, quantitative time series contained in previous

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

laboratory studies and vital signs are harder to process and likely less salient. Finally, while some prior diagnoses (e.g., diabetes, prior blockage) and procedures (e.g., prior stenting) relevant to blockages may be salient, these categories are far broader, including hundreds of other types of information that we also expect to be less salient.

Column (3) shows how these risk predictors correlate with the testing decision. Even after including risk from all other variable subsets, risk from symptoms stays positive (i.e., overweighted), as is risk from demographic information: a patient in the top quartile of symptom risk is 5.26 percentage points more likely to be tested, relative to other patients, and 0.78 percentage points for demographic risk.[46] This is equivalent to a patient moving from the 50th percentile of true (full) risk to the 89th and 62nd percentile, respectively. Prior quantitative information from laboratory studies and vital signs, though, has a negative sign, suggesting that physicians underweight or neglect this information. Finally, diagnoses are slightly overweighted while procedures are slightly underweighted. Taken together, these results are generally supportive of the salience model: risk signals from clearly salient inputs—demographics and symptoms—are attended to more than they should be, while more complex, less salient information—past quantitative vital signs and labs—are neglected.

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

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

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

This model has a crisp empirical prediction: at the same predicted risk, patients with more (less) representative symptoms are more (less) likely to be tested. We investigate this by first

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

identifying the set of symptoms that are potentially representative of blockage. To make this list, we identify those tested patients ultimately found to have blockages after testing and look back at their presenting symptom (limiting to 16 symptoms with frequency over 0.5% in this population; see Online Appendix Table A.16). For each symptom M, we calculate its representativeness for blockage: $\frac{Pr(M=1|B=1)}{Pr(M=1|B=0)}$. Nine symptoms have a ratio over 1, which we consider representative of blockage. Some are very common in the general population (e.g., chest pain, shortness of breath) and others are quite rare (e.g., presenting to the ER after a referral for a concern of possible blockage or because they were found unresponsive or in cardiac arrest by paramedics). The remaining seven symptoms are more common in the general population than in those with blockage (e.g., dizziness, nausea).

This allows us to build yet another risk predictor, restricting to representative symptoms. Table VI, column (4) shows the results of adding this to the regression we described previously (column (2)) with the predictor formed from all symptoms. With representative symptoms included, the all-symptom-based predictor becomes small and insignificant. The coefficient on the representative symptom-based predictor in column (4) is nearly double the magnitude of the all-symptom-based predictor in column (3).[47] This argues that while symptoms as a whole may be salient, representative symptoms drive physicians to test far more: they effectively cue the physician's mind to consider blockage. This effect is quantitatively large: the 7% in the highest quintile of representative symptom risk are 16.2 percentage points more likely to be tested, corresponding to an increase from the 50th to the 98th percentile of true risk.

Further, as shown in Online Appendix Figure A.14, patients whose risk comes disproportionately from representative symptoms (i.e., large $[\widehat{m}_{\text{represent}}(X_{ij}) - \widehat{m}(X_{ij})]$) are overrepresented in testing errors. Those in the top quintile of representativeness risk (relative to true risk) make up 34.3% of the low-risk tested; while the bottom quintile makes up 99.4% of the high-risk untested.[48]

47. Online Appendix Table A.18 confirms this new predictor has no incremental value for predicting yield.

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

V.C. *Implications for Incentive Policies*

The simultaneous presence of over- and underuse suggests that simple views of health care like "less is more" or "more is more" are insufficiently nuanced. Our results thus add to the growing body of work in health economics arguing for richer models of physician behavior (Kolstad 2013; Abaluck et al. 2016; Chandra and Staiger 2020; Chan, Gentzkow, and Yu 2022). Policy makers have long viewed health care through the lens of misaligned incentives that make physicians too eager to test. Implicit in this model is that physicians estimate risk correctly but simply set too low a threshold. This "less is more" model, which suggests that high-testing providers are wasteful relative to low-testing ones, has a clear practical implication that drives much of health policy in the United States and internationally: create incentives to test less, for example, via reimbursement schemes or capacity constraints. Yet our finding of systematic biases by physicians calls this approach into question: if physicians mispredict risk, incentives to cut care may do harm as well as good.

We empirically examine these potentially perverse consequences by asking, when physicians test less, which tests do they cut? The view of traditional models—and the hope of health policy—is that they cut the low-value tests. The top panel of Figure VII shows that this is not the case. Here we graph the probability of testing against predicted risk separately for each of the testing quartiles in our quasi-experiment (using the random effects model described above). Low-testing shifts do cut back on low-value tests: the lowest-risk patients are tested only 0.4% of the time, versus 3.0% on the highest-testing shifts. But they also cut back on high-value tests: the highest-risk patients are tested 5.8% of the time, versus 32.3% on the highest-testing shifts. In absolute terms, high-value tests suffer the biggest decline—26.5% fewer in low- versus high-testing regimes. In relative terms, low- and high-value tests fall by similar amounts: 87% versus 82%, respectively. In other words, less testing means less testing for everyone, regardless of risk. The bottom panel replicates these results in our nationally representative Medicare sample, where we sort hospitals into quintiles based on their testing rate, and again graph testing versus predicted risk for each quintile. We see the same result: hospitals that test more test everyone more.[49]

49. This exercise uses hospital referral regions to group hospitals, mirroring a large health policy literature that makes such cross-sectional comparisons.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

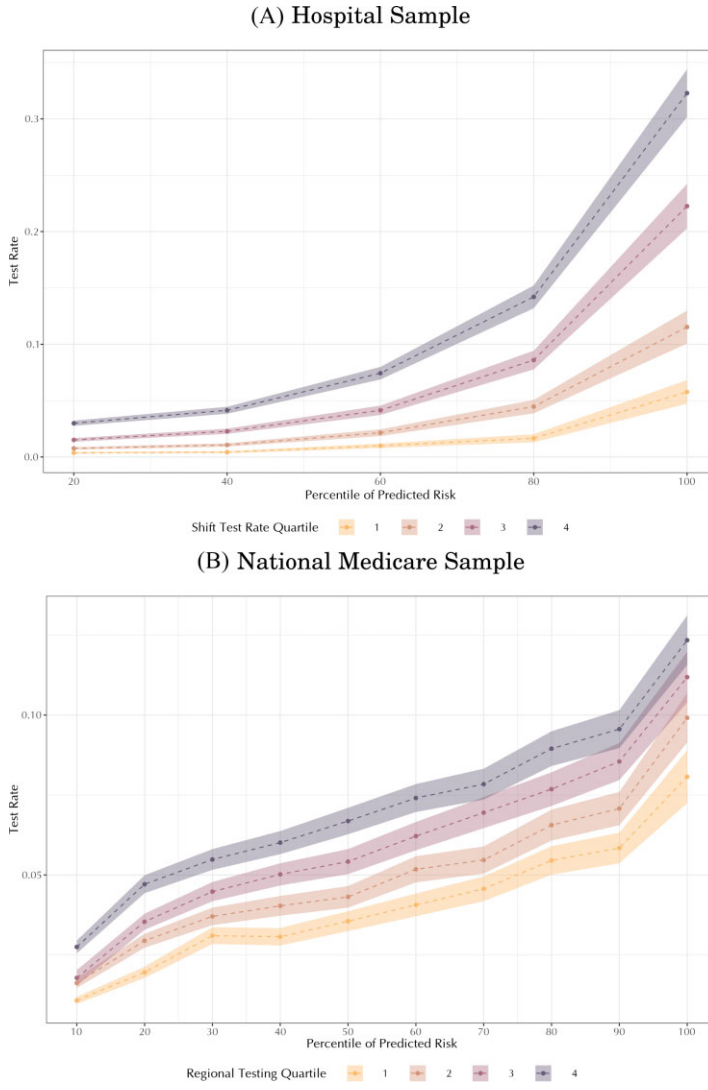

FIGURE VII

Variation in Testing Rates by Predicted Risk

Panel A shows variation in testing rates by predicted risk, in our quasi-experiment where patients are tested at higher or lower rates based on the triage staff working when they arrive. Panel B shows variation in testing rate by predicted risk, across all hospitals in the United States. Hospitals are binned into quartiles based on the overall testing rate of the hospital referral region in which they are located, to mirror cross-sectional analyses in the literature.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

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

V.D. The Role of Physician Experience

If incentives do not reduce inefficiency, what does? A natural candidate is physician experience, which we observe in our data. Though we cannot causally identify the effect of experience, correlations can be suggestive. In particular, we study how the correlation between physician decisions and patient risk varies with physician experience (as measured by years since residency). In Table VI, we regress testing on predicted risk, experience, and an interaction term between experience and patient risk. Column (5) shows that more experienced physicians test less on average: 1.68 percentage points or 0.05% for every year since residency. At the same time, experienced physicians are better able to match testing decisions to risk: with every year of experience, they test the lowest-risk patients 0.04 percentage points (2.81%) less, and the highest-risk 0.58 percentage points (1.06%) more.[50] These correlations provide suggestive evidence that physicians may learn over time, becoming more accurate with experience.

The results on experience in this section and the results on high- versus low-testing regimes tell distinct stories. On one hand, experienced physicians both test less and are more accurate. This

Naturally, these comparisons can be confounded. Although we lack the data to replicate the shift variation experiment, we do have an (albeit weaker) alternative, described in Online Appendix A.8.C. Testing typically requires an overnight stay after ED visits, but since hospital staffing is limited on weekends, patients who come in the day before a weekend are tested less. Online Appendix Figure A.10 shows that these reductions in testing reduce testing for all patients, irrespective of their actual risk.

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

echoes Chan, Gentzkow, and Yu (2022), who show a negative relationship between skill and testing levels. On the other hand, in Section V.C, we saw that less testing was uncorrelated with accuracy: testing fell across the risk distribution, including high-risk patients. This suggests that the relationship between testing level and accuracy is complex, and that care is needed to characterize it accurately. Understanding what leads physicians to be more or less accurate—and how that relates to testing level—is an important and open question.

VI. Conclusions

Much of machine learning applied to health care focuses on building tools to aid or substitute for humans: for example, algorithms that can match radiologists' performance on X-rays. Our work suggests a very different use for machine learning in health care: as a tool to understand humans and the health systems they work in.

This approach allows us to precisely characterize inefficiencies. Current empirical approaches in health policy rely largely on aggregates: for example, do tests on average yield enough positives to justify their costs (Weinstein et al. 1996; Sanders et al. 2016)? By that metric, testing appears highly efficient, at only $89,714 per life-year in our data. The granularity of algorithmic predictions, by contrast, reveals both under- and overuse. This reframes the discussion away from how many people get tested—too many, or too few?—to one about who gets tested. In a very conservative simulation of optimal testing, total testing would drop by 47%, but the composition of the tested would change radically: 29% of efficient tests would be new, in patients physicians do not currently test; and tests would go from costing $89,714 to $59,390 per life-year. The importance of composition in turn calls into question the central role of incentives in policy. By changing the level of testing alone, they may improve one inefficiency (overuse) while aggravating another (underuse).

Despite the great promise of algorithms for diagnosing and improving human inefficiencies, great care is needed when comparing human decisions and algorithmic predictions. As we saw, when physician and algorithm disagree, we cannot just assume the algorithm is correct: unobserved variables confound algorithmic predictions. This selection bias pervades machine learning applications in health and elsewhere, appearing whenever

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

algorithms are trained on data produced by the humans they seek to influence.[51] Once acknowledged, we show these problems can be tackled: by developing new labels grounded in domain expertise, and via quasi-experimental methods from the causal inference toolkit. But ignoring this bias risks stacking the deck in favor of algorithms: assuming away physician private information means algorithms can, by construction, never do worse than the human—a misleading comparison.

Finally, our findings suggest a role for algorithmic predictions in interventions to increase efficiency. Most obviously, because they are built on EHR data, our predictions can be delivered to physicians in real time. Rather than replacing their judgment, they can be combined with physician private information. At the payment level, a system of precision pricing could tie incentives and reimbursements for testing to patient-level predicted risk and testing outcomes. Or predictions could be used as an educational tool, during physician training or as continuing medical education. We found accuracy improves with experience, but using algorithms to hasten the learning process would be valuable: human trial and error is a costly way to learn in medicine.

UNIVERSITY OF CHICAGO, UNITED STATES
UNIVERSITY OF CALIFORNIA, BERKELEY, UNITED STATES

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online.

DATA AVAILABILITY

Code replicating the tables and figures in this article can be found in Mullainathan and Obermeyer (2021) in the Harvard Dataverse, https://doi.org/10.7910/DVN/IUMIO6.

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

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

References

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

Ægisdóttir, Stefanía, Michael J. White, Paul M. Spengler, Alan S. Maugherman, Linda A. Anderson, Robert S. Cook, Cassandra N. Nichols, and Georgios K. Lampropoulos et al., "The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction," *Counseling Psychologist*, 34 (2006), 341–382.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb, "Exploring the Impact of Artificial Intelligence: Prediction Versus Judgment," *Information Economics and Policy*, 47 (2019), 1–6.

Al-Lamee, Rasha, David Thompson, Hakim-Moulay Dehbi, Sayan Sen, Kare Tang, John Davies, Thomas Keeble, and Michael Mielewczik et al., "Percutaneous Coronary Intervention in Stable Angina (ORBITA): A Double-Blind, Randomised Controlled Trial," *Lancet*, 391 (2018), 31–40.

Amsterdam, Ezra A., Nanette K. Wenger, Ralph G. Brindis, Donald E. Casey, Theodore G. Ganiats, David R. Holmes, Allan S. Jaffe, and Hani Jneid et al., "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," *Journal of the American College of Cardiology*, 64 (2014), e139–e228.

Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 53 (1963), 941–973.

Backus, Barbra E., A. Jacob Six, Johannes C. Kelder, Thomas P. Mast, Frederieke van den Akker, E. Gijis Mast, Stefan HJ Monnink, and Rob M. van Tooren et al., "Chest Pain in the Emergency Room: A Multicenter Validation of the HEART Score," *Critical Pathways in Cardiology*, 9 (2010), 164–169.

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein, "Behavioral Hazard in Health Insurance," *Quarterly Journal of Economics*, 130 (2015), 1623–1667.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes," *Quarterly Journal of Economics*, 131 (2016), 1753–1794.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience Theory of Choice under Risk," *Quarterly Journal of Economics*, 127 (2012), 1243–1285.

———, "Memory, Attention, and Choice," *Quarterly Journal of Economics*, 135 (2020), 1399–1442.

Brot-Goldberg, Zarek C., Amitabh Chandra, Benjamin R. Handel, and Jonathan T. Kolstad, "What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics," *Quarterly Journal of Economics*, 132 (2017), 1261–1318.

Camerer, Colin, "Artificial Intelligence and Behavioral Economics," in *The Economics of Artificial Intelligence: An Agenda*, (Chicago: University of Chicago Press, 2019), 587–610.

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

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

Chandra, Amitabh, Evan Flack, and Ziad Obermeyer, "The Health Costs of Cost-Sharing," NBER Working Paper no. 28439, 2021.

Chandra, Amitabh, Jonathan Gruber, and Robin McKnight, "Patient Cost-Sharing and Hospitalization Offsets in the Elderly," *American Economic Review*, 100 (2010), 193–213.

Chandra, Amitabh, and Douglas O. Staiger, "Identifying Sources of Inefficiency in Healthcare," *Quarterly Journal of Economics*, 135 (2020), 785–843.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

Croskerry, Pat, "Achieving Quality in Clinical Decision Making: Cognitive Strategies and Detection of Bias," *Academic Emergency Medicine*, 9 (2002), 1184–1204.
Dawes, R. M., D. Faust, and P. E. Meehl, "Clinical versus Actuarial Judgment," *Science*, 243 (1989), 1668–1674.
Elstein, Arthur S., "Heuristics and Biases: Selected Errors in Clinical Reasoning," *Academic Medicine*, 74 (1999), 791–794.
Fisher, Elliott S., David E. Wennberg, Thérèse A. Stukel, Daniel J. Gottlieb, F. L. Lucas, and Étoile L. Pinder, "The Implications of Regional Variations in Medicare Spending. Part 2: Health Outcomes and Satisfaction with Care," *Annals of Internal Medicine*, 138 (2003), 288–298.
Gabaix, Xavier, "A Sparsity-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 129 (2014), 1661–1710.
———, "Behavioral Inattention," in *Handbook of Behavioral Economics, Foundations and Applications*, vol. 2, (Amsterdam: Elsevier, 2019), 261–343.
Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri, "Economic Predictions with Big Data: The Illusion of Sparsity," *Econometrica*, 89 (2021), 2409–2437.
Graber, Mark L., Nancy Franklin, and Ruthanna Gordon, "Diagnostic Error in Internal Medicine," *Archives of Internal Medicine*, 165 (2005), 1493–1499.
Hamon, Martial, Jean-Claude Baron, Fausto Viader, and Michéle Hamon, "Periprocedural Stroke and Cardiac Catheterization," *Circulation*, 118 (2008), 678–683.
Handel, Benjamin R., and Jonathan T. Kolstad, "Health Insurance for 'Humans': Information Frictions, Plan Choice, and Consumer Welfare," *American Economic Review*, 105 (2015), 2449–2500.
IOM (Institute of Medicine), *Improving Diagnosis in Health Care* (Washington, DC: National Academies Press, 2015).
Kahneman, Daniel, and Amos Tversky, "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology*, 3 (1972), 430–454.
Kallus, Nathan, and Angela Zhou, "Confounding-Robust Policy Improvement," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18 (Red Hook, NY: Curran Associates Inc., 2018), 9289–9299.
Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 133 (2018), 237–293.
Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer, "Prediction Policy Problems," *American Economic Review*, 105 (2015), 491–495.
Kolstad, Jonathan T., "Information and Quality When Motivation Is Intrinsic: Evidence from Surgeon Report Cards," *American Economic Review*, 103 (2013), 2875–2910.
Mahoney, Elizabeth M., Claudine T. Jurkovitz, Haitao Chu, Edmund R. Becker, Steven Culler, Andrzej S. Kosinski, Debbie H. Robertson, and Charles Alexander et al., "Cost and Cost-Effectiveness of an Early Invasive vs Conservative Strategy for the Treatment of Unstable Angina and Non–ST-Segment Elevation Myocardial Infarction," *Journal of the American Medical Association*, 288 (2002), 1851–1858.
Mullainathan, Sendhil, "A Memory-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 117 (2002), 735–774.
Mullainathan, Sendhil, and Ziad Obermeyer, "Replication Data for: 'Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care'," (2021), Harvard Dataverse, https://doi.org/10.7910/DVN/IUMIO6.
Mullainathan, Sendhil, and Jann Spiess, "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31 (2017), 87–106.
Norris, Samuel, "Examiner Inconsistency: Evidence from Refugee Appeals," Becker Friedman Institute Working Paper, 2019.

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025

Obermeyer, Ziad, Brent Cohn, Michael Wilson, Anupam B. Jena, and David M. Cutler, "Early Death after Discharge from Emergency Departments: Analysis of National US Insurance Claims Data," *British Medical Journal*, 356 (2017), https://doi.org/10.1136/bmj.j239.

Pauly, Mark V., "The Economics of Moral Hazard: Comment," *American Economic Review*, 58 (1968), 531–537.

Paxton, Chris, Alexandru Niculescu-Mizil, and Suchi Saria, "Developing Predictive Models Using Electronic Medical Records: Challenges and Pitfalls," in *AMIA Annual Symposium Proceedings*, vol. 2013 (Bethesda, MD: American Medical Informatics Association, 2013), 1109–1115.

Pope, J. Hector, Tom P. Aufderheide, Robin Ruthazer, Robert H. Woolard, James A. Feldman, Joni R. Beshansky, John L. Griffith, and Harry P. Selker, "Missed Diagnoses of Acute Cardiac Ischemia in the Emergency Department," *New England Journal of Medicine*, 342 (2000), 1163–1170.

Prasad, Vinay, Michael Cheung, and Adam Cifu, "Chest Pain in the Emergency Department: The Case against Our Current Practice of Routine Noninvasive Testing," *Archives of Internal Medicine*, 172 (2012), 1506–1509.

Rambachan, Ashesh, "Identifying Prediction Mistakes in Observational Data," Working Paper, 2021.

Redelmeier, Donald A., Lorraine E. Ferris, Jack V. Tu, Janet E. Hux, and Michael J. Schull, "Problems for Clinical Judgement: Introducing Cognitive Psychology As One More Basic Science," *Canadian Medical Association Journal*, 164 (2001), 358–364.

Sanders, Gillian D., Peter J. Neumann, Anirban Basu, Dan W. Brock, David Feeny, Murray Krahn, Karen M. Kuntz, and David O. Meltzer et al., "Recommendations for Conduct, Methodological Practices, and Reporting of Cost-Effectiveness Analyses: Second Panel on Cost-Effectiveness in Health and Medicine," *Journal of the American Medical Association*, 316 (2016), 1093–1103.

Shanmugam, Vimalraj Bogana, Richard Harper, Ian Meredith, Yuvaraj Malaiapan, and Peter J. Psaltis, "An Overview of PCI in the Very Elderly," *Journal of Geriatric Cardiology*, 12 (2015), 174–184.

Simon, Herbert A., "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics*, 69 (1955), 99–118.

Sims, Christopher A., "Implications of Rational Inattention," *Journal of Monetary Economics*, 50 (2003), 665–690.

Tversky, Amos, and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science*, 185 (1974), 1124–1131.

Wei, Wei-Qi, Qiping Feng, Peter Weeke, William Bush, Magarya S. Waitara, Otito F. Iwuchukwu, Dan M. Roden, and Russell A. Wilke et al., "Creation and Validation of an EMR-based Algorithm for Identifying Major Adverse Cardiac Events while on Statins," *AMIA Summits on Translational Science Proceedings*, 2014 (2014), 112–119.

Weinstein, Milton C., Louise B. Russell, Marthe R. Gold, and Joanna E. Siegel (eds.), *Cost-Effectiveness in Health and Medicine* (Oxford: Oxford University Press, 1996).

Downloaded from https://academic.oup.com/qje/article/137/2/679/6449024 by MIT Libraries user on 23 July 2025