Appendix to: Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care

Sendhil Mullainathan, Ziad Obermeyer

October 2021

Contents

1 Additional Details on Sample and Key Variables

1.1 Hospital Cohort

For our hospital electronic health record dataset, we begin with 326,126 ED visits (indexed by i), by 150,862 patients (indexed by j), over a five-and-a-half-year period from January 2010 through May 2015. We exclude visits in which the patient died in the ED and thus before they could be tested (0.07%); visits preceded by recent known heart attack or its treatment (e.g., catheterization, stenting in the 30 days prior to ED visits), for whom testing may represent follow up of a known problem, rather than diagnosis of a new one (0.2%); and patients with contraindications $K = 1$ to invasive treatment for heart attack, due to general poor health (patients 80 years of age or older: 7.4%; those with poor-prognosis conditions diagnosed in the year prior, e.g., known metastatic cancer, dementia, hospice or nursing home care, etc.: 16.6%).

1.2 Construction of Key Variables

A major but under-appreciated challenge in working with claims and electronic health record data is accurate measurement of clinical tests and outcomes. A straightforward concept like 'stress test' or 'cardiac catheterization' is represented in a range of evolving procedure codes and test results. There is no straightforward way to capture these: for example, widely cited papers on testing for heart attack use partially non-overlapping sets of 20–30 codes to identify procedures (e.g., Sheffield et al., 2013 vs. Schwartz et al., 2014 vs. Shreibati, Baker, and Hlatky, 2011). The most commonly used procedure coding system (Current Procedural Terminology, adapted for use with Medicare claims as the Healthcare Common Procedure Coding System) is modified every year, with significant changes that, in our data, led to major discontinuities in testing rates for the same hospital over time as codes and coding practices changed. To deal with this, we performed a comprehensive search of the literature as well as these coding databases. We ultimately identified 59 distinct codes for catheterization and 106 for stress test (detailed in Supplement: Billing Codes). Using our national Medicare claims dataset for generalizability, we find that relative to those typically used in the literature, these additional codes added 11% of tests and 5% of interventions.

Category	Type	Codes
Tests		
Stress Test	HCPCS	75559, 75560, 75561, 75562, 75563, 75564, 75571, 75572, 75573, 75574, 78402, 78403, 78404, 78407, 78411, 78412, 78414, 78415, 78418, 78419, 78420, 78422, 78424, 78428, 78435, 78451, 78452, 78453, 78454, 78459, 78460, 78461, 78462, 78463, 78464, 78465, 78466, 78467, 78468, 78469, 78470, 78471, 78472, 78473, 78474, 78475, 78476, 78477, 78478, 78479, 78480, 78481, 78483, 78484, 78485, 78486, 78487, 78489, 78491, 78492, 78494, 93015, 93016, 93017, 93018, 93024, 93350, 93351, 93352, 0144T, 0145T, 0146T, 0147T, 0148T, 0149T, 0151T, C8928, C8930, G0030, G0031, G0032, G0033, G0034, G0035, G0036, G0037, G0038, G0039, G0040, G0041, G0042, G0043, G0044, G0045, G0046, G0047, G8961, G8962, G8963, G8964, G8965, G8966

	ICD-9	8941, 8942, 8943, 8944
Catheterization	HCPCS	75523, 75524, 75527, 75528, 93452, 93453, 93454, 93455, 93456, 93457, 93458, 93459, 93460, 93461, 93462, 93508, 93510, 93511, 93514, 93524, 93526, 93527, 93528, 93529, 93539, 93540, 93541, 93542, 93543, 93545, 93546, 93547, 93548, 93549, 93550, 93551, 93552, 93553, 93555, 93556, 93563, 93564, 93565, 93566, 93567, 93568
	ICD-9	3722, 3723, 24, 66, 3601, 3602, 3603, 3604, 3605, 3606, 3607, 3608, 3609

Revascularization

Stent	HCPCS	92920, 92921, 92924, 92925, 92928, 92929, 92933, 92934, 92937, 92938, 92941, 92941, 92975, 92977, 92980, 92981, 92982, 92984, 92995, 92996, 92997, 92998, C9600, C9601, C9602, C9603, C9604, C9605, C9605, C9606, G0290, G0291
	ICD-9	3601, 3602, 3603, 3604, 3605, 3606, 3607, 3608, 3609, 0066
Thrombolysis	HCPCS	J3101, J2997, J2995, J2993
	ICD-9	991
CABG	HCPCS	33510, 33511, 33512, 33513, 33514, 33516, 33517, 33518, 33519, 33520, 33521, 33522, 33523, 33525, 33528, 33530, 33533, 33534, 33535, 33536, 33560, 33570, 33572, 33575, 35500, 35600, 93351, 4110F, S2205, S2206, S2207, S2208, S2209
	ICD-9	3610, 3611, 3612, 3613, 3614, 3615, 3616, 3617, 3618, 3619

Adverse events

Heart attack	ICD-9	410, 4100, 41000, 41001, 41002, 41003, 41004, 41005, 41006, 41007, 41008, 41009, 41010, 41011, 41012, 41013, 41014, 41015, 41016, 41017, 41018, 41019, 41020, 41021, 41022, 41023, 41024, 41025, 41026, 41027, 41028, 41029, 41030, 41031, 41032, 41033, 41034, 41035, 41036, 41037, 41038, 41039, 41040, 41041, 41042, 41043, 41044, 41045, 41046, 41047, 41048, 41049, 41050, 41051, 41052, 41053, 41054, 41055, 41056, 41057, 41058, 41059, 41060, 41061, 41062, 41063, 41064, 41065, 41066, 41067, 41068, 41069, 41070, 41071, 41072, 41073, 41074, 41075, 41076, 41077, 41078, 41079, 41080, 41081, 41082, 41083, 41084, 41085, 41086, 41087, 41088, 41089, 41090, 41091, 41092, 41093, 41094, 41095, 41096, 41097, 41098, 41099, 411, 4110, 4111, 41100, 41101, 41102, 41103, 41104, 41105, 41106, 41107, 41108, 41109, 41110, 41111, 41112, 41113, 41114, 41115, 41116, 41117, 41118, 41119, 41120, 41121, 41122, 41123, 41124, 41125, 41126, 41127, 41128, 41129, 41130, 41131, 41132, 41133, 41134, 41135, 41136, 41137, 41138, 41139, 41140, 41141, 41142, 41143, 41144, 41145, 41146, 41147, 41148, 41149, 41150, 41151, 41152, 41153, 41154, 41155, 41156, 41157, 41158, 41159, 41160, 41161, 41162, 41163, 41164, 41165, 41166, 41167, 41168, 41169, 41170, 41171, 41172, 41173, 41174, 41175, 41176, 41177, 41178, 41179, 41180, 41181, 41182, 41183, 41184, 41185, 41186, 41187, 41188, 41189, 41190, 41191, 41192, 41193, 41194, 41195, 41196, 41197, 41198, 41199
Cardiac arrest	HCPCS	92950
	ICD-9	4275, 4274, 42741, 42742, 78551

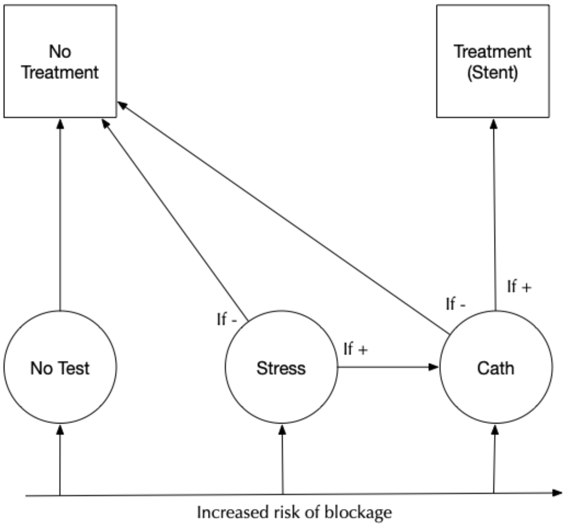
Table A.1: ICD-9 and HCPCS for testing, revascularization, and adverse events.

1.3 Key Assumptions

Time Window We use a 10-day window after visits to determine if a given patient is tested and treated. We chose this based on guidelines for testing, which range from, e.g., 72 hours in Amsterdam et al. (2014) to 1-2 weeks in Brown et al. (2018). Practically, in our sample, most tests (81%) are done either during the ED visit or in the 72 hours after (often after an overnight stay in the Observation Unit from the ED visit). The longer window accounts for the minority of tests and treatments done largely in the course of an expedited outpatient referral to a primary care provider or a cardiologist.

Testing By 'testing for heart attack' we mean testing for an acute blockage obstructing blood flow through one or more coronary arteries. We group two types of test for blockage together to define testing $T_{ij} = 1$: stress testing and catheterization. We do so because, practically, the basic intent of both tests is to diagnose these blockages. But there are also important nuances of these two testing types relevant to our subsequent analyses, that we delve into here.

First and most straightforward is catheterization, an invasive procedure that is both the definitive test, and also the means by which treatment in the form of stenting is delivered. Catheterization can either be done as the initial test, or the physician can choose to start with a stress test: a set of non-invasive procedures like putting the patient on a treadmill with electrocardiographic monitoring to look for signs of blockage, or radiological studies that visually quantify potential blockages. Importantly, while stress tests can suggest blockages, they cannot confirm or treat them. So if the physician chooses to first perform a stress test, it functions only as a first, lower-cost step to screen out negatives. If positive, the patient must proceed to catheterization for confirmation and definitive treatment.



The costs and benefits of these two categories of tests are different. Because it is an invasive procedure, the cost of catheterization is high. Financial costs total nearly $30,000, and there is a small but measurable set of procedural risks (most catastrophically, stroke). The benefit, of course, comes in the form of treatment for heart attack, if the catheterization identifies a blockage. Stress tests, because they are non-invasive, have lower costs: financial costs are around $4,000, and the health costs are minimal. Of course, if they come back positive, the patient will go on to pay the costs of catheterization, as well as the cost of the stress test. The benefit of stress tests is thus also lower: because they are not perfect (i.e., the probability they will be positive is less than 1), the joint probability of both a stress test and a catheterization will be positive is less than the probability that catheterization alone would be positive. So while they lead to the same benefit when both tests are positive, via delivery of treatment for heart attack, they are less likely to lead to benefit on average. As a result of this calculus, summarized in the diagram above, it is intuitive and efficient for physicians to begin with stress testing for lower-risk patients: their expected benefit—the likelihood of testing positive, going on to catheterization, and receiving beneficial treatment—is less than the cost of catheterization, but greater than the cost of the stress test. Higher-risk patients, on the other hand, should go straight to catheterization without incurring the additional cost and likelihood of false negative from stress testing.

Our analysis accounts for these nuances in two ways. First, because physicians use both types of test to answer the same question—does their patient have an acute coronary blockage—we group them together in our definition of testing T_{ij}. (By contrast, in other settings, largely outpatient clinics, these tests may be done for a variety of reasons: to characterize a baseline for future reference, to plan an elective surgery, etc.) Importantly, we allow the financial costs of testing to vary in our cost-benefit analysis, and thus keep careful track of which type of test is done on which patient. Overall, we identified 7,320 tested visits, and 242,343 untested visits (which are described in more detail below). Of the tested, 4,876 had stress tests, and 3,080 had cardiac catheterization; 636 had both a stress test and subsequent catheterization, indicating a positive stress test.

Second, the existence of different testing strategies means there is more than one counterfactual we might wish to consider in our analyses of testing decisions. For our main analyses, we consider a specific counterfactual: eliminating (or adding) the specific kind of test the physician chose to do in our dataset, in a patient with a given level of predicted risk. Other counterfactuals are also of interest, for example, eliminating all stress tests, or all catheterizations. We might also be interested in counterfactuals where we replace all catheterizations with stress tests, and vice-versa. We explore some of these in Appendix 3, where we first calculate the value of testing with catheterization or stress test alone, to simulate a scenario where one type of test is eliminated (added). We cannot fully simulate substitutions between testing types: for example, we do not know what stress tests would have shown in catheterized patients. But we can calculate a conservative bound on substitution of catheterization for stress tests, where we drop only negative stress tests (those that do not progress to catheterization). This simulates a near-perfect strategy of replacing current stress tests with catheterization.

Treatment By 'treatment for heart attack,' we mean invasive procedures performed to open acute blockages in the coronary arteries. We group two types of treatments for blockage together to define treatment $S_{ij} = 1$: stenting (percutaneous intervention: PCI) and open-heart surgery (coronary artery bypass grafting, CABG). 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. Stenting is far more common, while bypass surgery (CABG) is done in more severe cases.

We did not include intravenous thrombolysis. Previously the standard of care, this treatment involves administration of a clot-busting drug (thrombolytic) that dissolves clots—but does so everywhere in the body, not just in the heart. Thrombolysis is not performed at the hospital we study (which has a catheterization facility on-site). In our analyses of Medicare data as well, we decided not to include it for two reasons. First, for comparability to our hospital-based result, and second because it is not captured nearly as reliably: it is the administration of a medication, not a procedure (Kleindorfer Dawn et al., 2008). In addition, it is used for other purposes besides treating heart attack (e.g., stroke). This means we would under-capture treatments at some hospitals (especially smaller rural ones), which continue to use thrombolysis. While this practice fell out of favor in the 1990s with the advent of stenting, not all hospitals have a catheterization facility. And while growing evidence has pushed most hospitals to transfer patients to a hospital that does, because it is superior to thrombolysis despite the delays induced by transport time (Widimský et al., 2003), some remote hospitals will offer thrombolysis when transit times are particularly long.

Yield of Testing To define test yield Y_{ij}, we assume that a positive test always leads to treatment and thus set $Y_{ij} = S_{ij}$ for the tested. The basis for this assumption is two-fold. First, catheterization is physically required to deliver a stent to the location of a blockage—indeed, it is the same physical procedure. The less common treatment option, open-heart surgery, likewise requires prior catheteriztion, to determine eligibility for surgery, and to map out arterial anatomy in preparation for surgery. Second, it would be ethically dubious for a cardiologist to subject a patient to the risks of emergency catheterization unless she has already decided the patient would benefit from treatment if a blockage is detected.

We address one potential problem with this assumption in the main text: if physicians over-treat conditional on test results (e.g., because of moral hazard, or false-positive tests). This is not a problem for our analyses in two ways. First, our estimates of over-testing are based on tests that do not lead to treatment as low-value. If some tests that *do* lead to treatment are in fact low-value, this could cause us to under-state the extent of over-testing. Practically, this means that our estimates of over-testing are a lower bound. Second, in our cost-effectiveness analysis, we use treatment effects from real-world trials that include these false positives. Because we do not use test results to quantify under-testing in any way, instead relying on adverse event rates and a natural experiment, our estimates are unaffected by this bias.

Beyond over-treatment, we are not aware of evidence of other variation in treatment conditional on these particular test results. Test results for catheterization are based on an objective measure of blood flow through the coronary arteries, which either visualizes a blockage or does not; based on these results, either a stent is placed in the blockage or not. So while there is ample evidence of bias in testing decisions—for example, doctors are less likely to refer patients for testing for heart attack when presented with vignettes accompanied by randomly assigned pictures of women and minorities (Schulman et al., 1999)—bias in treatment, conditional on test results, is likely to be less widespread (though not impossible). A physician would need to overrule the objective measures, and clinical guidelines, to deviate. There is no clear reason to believe that there are widespread types of biases that would allow testing, but discourage treatment in tested patients.

Finally, we might worry about correlation between the testing threshold and the treatment decision. This is mitigated somewhat by the fact that the physician referring the patient for testing (the emergency physician) and the physician performing the test and treatment procedure (the cardiologist) are two different people.

Adverse Events Our measure of adverse events resulting from untreated (and undiagnosed) heart attack is drawn from the clinical literature. Because effective treatment for heart attack emerged only in the early 1980s, there is a large body of fairly recent research documenting the fate of patients with untreated heart attack.[1] Clinical trials of diagnostic and treatment interventions for heart attack (e.g., CT-angiography to diagnose coronary disease, e.g., Litt et al. (2012); or statins to treat high lipids, e.g., Ridker et al. (2008)) commonly use a basket of events derived from this literature, 'major adverse cardiac events,' as their primary outcome. We replicate this in our data, using methods similar to observational clinical studies (e.g., for decision rules: Than et al. (2011), Poldervaart et al. (2017), and Sharp, Broder, and Sun (2018)) that have shown excellent agreement with expert judgment after chart review (e.g., Wei et al. (2014)).

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

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

patient can reach the hospital in time to be diagnosed or treated, the patient might simply drop dead—often outside of the hospital. This normally poses a major problem for observational studies, because out-of-hospital deaths are not recorded in hospital records. To deal with this, we link hospital records to state Social Security data. Together, these data allow us to form A_{ij}, our proxy for blockage in the untested.

To use the rate of these adverse events as a measure of under-testing, we must compare it to some upper bound. This seems complex, because even in a short time window after visit j, not all adverse events in untreated patients result from heart attacks at the time of the visit: in temrs of our model, we would like to know $E[\frac{A_{ij}-\mu}{\zeta}]$, but do not observe μ or ζ. So we instead pin down the adverse event threshold in terms of the *total* adverse event rate in untreated patients, including the base rate. To do so, we draw on the clinical decision rule literature, in particular those used to allocate tests for heart attack.

Several studies define a maximum allowable rate of adverse events in untested, and thus untreated, patients. An additional advantage of using these studies is that they are widely accepted by clinicians. Fortunately, such bounds are common in the clinical literature on decision rules, in particular those that seek to help doctors test for heart attack (e.g., TIMI: Antman et al. (2000), GRACE: Tang, Wong, and Herbison (2007), HEART: Backus et al. (2010) and subsequent validation studies, e.g., Than et al. (2011), Poldervaart et al. (2017), and Sharp, Broder, and Sun (2018)), as well as studies of new diagnostic technologies (e.g., CT-angiography: Litt et al. (2012)) or guidelines for preventative treatment (e.g., with statins: Ridker et al. (2008)) of heart attack. All these studies share the need to define a maximum allowable rate of adverse events in untested (and thus untreated) patients.

Practically, if a group of patients is ex post found to have an adverse event rate of over some bound, the recommendation is that this group should have been tested. This line of research guides routine testing and management decisions in clinics and hospitals, and underlies recommendations from professional societies. It thus gives us objective thresholds for levels of risk that would mandate testing (inclusive of the base rate of adverse events μ): 2% over the 30-day window after visits.[2] We do not assume this threshold is socially or physiologically optimal, only that it represents current physician understanding of who should be tested.

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

Finally, we ascertain mortality by linking EHR data to Social Security Death Index data. This

[2] As another point of comparison, surveys of emergency doctors ask about their willingness to accept a heart attack miss rate. These find a tolerance of up to 1% for heart attacks in the ER (Than et al., 2013).

is important because otherwise, we would only see mortality if the patient died in-hospital, and in particular at a hospital that forms part of the same health system we study for care. Linkage was performed via patient Social Security number, so may under-capture deaths in patients without SSNs.

2 Cost-effectiveness Analysis

2.1 Approach and Parameters

Our cost effectiveness analysis accounts for the costs of testing c_T and treatment c_S. We separately calculate the direct cost of both non-invasive tests like stress tests, c_N and invasive catheterization c_I, as well as the health costs due to testing.

We use a set of parameters listed in Table A.2 to calculate cost per quality adjusted life year. The direct costs of testing and treatment are fairly straightforward to measure, using standardized Medicare fee schedules.[3] For the health costs of testing, we focus on the risk of stroke during catheterization, based on the literature studying the individual complications of testing (in health care dollars and quality of life). We considered including several other costs, but decided not to: (i) we do not know of any credible estimates for the health costs of exposure to radiation, or (ii) whether the association between sudden death and treadmill testing is causal or simply reflects confounding; and we did not include costs of arterial pseudoaneurysm because of a combination of low rates and very low disability weights and direct costs identified on literature review.

The benefits, in terms of quality adjusted life years, are less straightforward to quantify. Our strategy follows the general approach laid out by Tan, Kuo, and Goodwin, 2013, and begins by estimating the life expectancy using a Cox proportional-hazard model. The model incorporates age, sex, and a variety of comorbidities which we are able to determine using patient history from EHR data. We take life expectancy to be the number of months after which the survival rate is approximatley 50%. Using this model, we are able to estimate expected life-years remaining for all the patients in our sample. From this we use standard assumptions regarding the probability of fatal vs. non-fatal heart attack to arrive at years lost due to heart attack. This number combines both death and disability, the latter captured using a standard discount rate that reflects the sequelae of heart attack—angina pectoris, heart failure, etc.

We then use estimates from the literature to estimate the fraction of these losses that would be averted by timely treatment, to arrive at a treatment effect at the individual level expressed in life years. To do so, we focus on randomized trial evidence describing the benefits of treatment of heart attack, reviewed extensively in Amsterdam et al., 2014. This review provides estimates from a range of randomized trials testing treatments for non-ST-elevation ACS (as opposed to the more severe ST-ACS); we chose this since we imagine most of the patients who undergoing testing have the less severe form of heart attack (of note, treatment effects are larger for ST-ACS). From this range, we chose the estimate from Bavry et al., 2006 as our primary estimate. This study is a meta-analysis of trials where patients with non-ST-ACS were randomized to either (i) early and universal treatment with stenting (typically within 2-3 days), vs. (ii) an "as needed" approach where patients were observed and treated medically (i.e., with medications, e.g., aspirin, statins, etc.) unless they deteriorated. We view this choice as quite conservative: this study compared interventions of the kind we study (i.e., stenting) to a counterfactual of watchful waiting with ongoing drug treatments, all in patients diagnosed with heart attacks, while the counterfactual in our study was no diagnosis and no treatment. We thus also show the range of possible treatment effects from the other studies in Amsterdam et al., 2014, which end up being both above and below the estimate of 0.25 for mortality at two years from Bavry et al., 2006. Of note, the upper bound is from a study by Mills et al., 2011, which quantifies the mortality benefit of a more sensitive diagnostic test in the ED; of

[3] Rates calculated from our sample would be affected by sample restrictions, and would not be standardized for labor and material costs across regions, so we opted to use published rates.

Table A.2: Cost-Effectiveness Parameters and References

	Value	Notes & References
Costs		
Non-invasive testing[1]	\$4000	CMS, 2016a; CMS, 2016b; Sun et al., 2014; Rydman et al., 1998; Eisenberg et al., 2005
Catheterization	\$28,000	
Stenting	\$15,000	
CABG	\$20,000	
Risks		
Stroke risk	0.0044	Hamon et al., 2008
Stroke disability weight	0.266	Hong and Saver, 2009
Lifetime direct cost of stroke [2]	\$170,000	Taylor et al., 1996
Heart attack		
p(fatal)	0.2	Mahoney et al., 2002
p(nonfatal)	0.8	
Nonfatal disability weight	0.125	
Life expectancy with heart disease		Peeters et al., 2002
Treatment effect		
Lower bound	0.20	Amsterdam et al., 2014; Bavry et al., 2006; Mills et al., 2011
Most likely	0.25	
Upper bound	0.30	

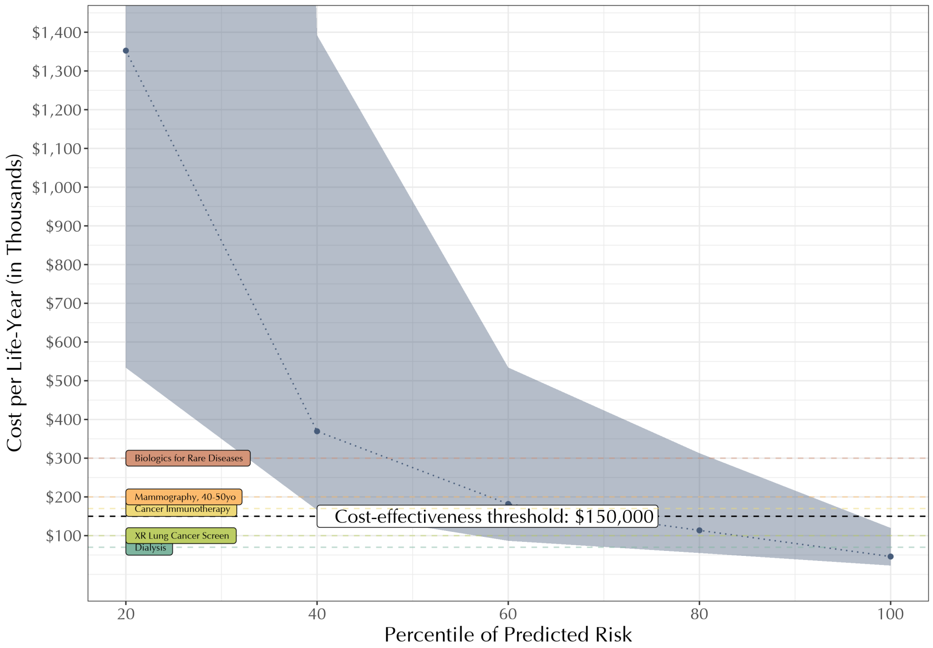
[1] Rates by Medicare Severity Diagnosis Related Groups calculated using the national adjusted full update with standardized labor, non-labor and capital amounts. Testing cost reflects direct costs (ranging from \$400 for treadmill testing to \$2,000 for imaging), physician services (\$500) and facility fees (ranging from \$1,000 for observation care to \$2,000 for inpatient stays, with the latter more common). [2] Includes only medical care; does not include lost productivity.

all the studies reviewed, this is perhaps closest to the counterfactual we attempt to study here. In the main text, we use 0.2 and 0.3 as our treatment effect lower and upper bounds, respectively. In Figure A.1, we provide analogous results using more relaxed bounds: 0.1 and 0.5.

2.2 Comparison of Clinical Guidelines for Testing Based on Adverse Events to Cost-Effectiveness Measures

In the main text, we discuss two different thresholds for determining whether or not a patient should be tested for heart attack: the 2% miss-rate guideline for the untested, and the \$150K cost-effectiveness threshold in the tested. Here, we use observed adverse outcome rates in the untested and observed cost-effectiveness of tests to relate these two metrics. Specifically, we divide the sample into ventiles of predicted risk, with ventile cutoffs determined in the tested. Then, for each ventile bin, we calculate the adverse event rate in the untested and the yield in the tested. In Figure A.2, we plot these rates as points, along with a best fit line. We then add a vertical line at $x = 0.120$, which is the yield rate associated with 150K cost-effectiveness. Finally, we add a horizontal line where the first two lines intersect: $y = 0.032$. We interpret this as the acceptable miss rate implied by the \$150K cost-effectiveness target. Indeed, it is statistically indistinguishable

Figure A.1: Cost-Effectiveness with 10-50% Treatment Effect Bounds



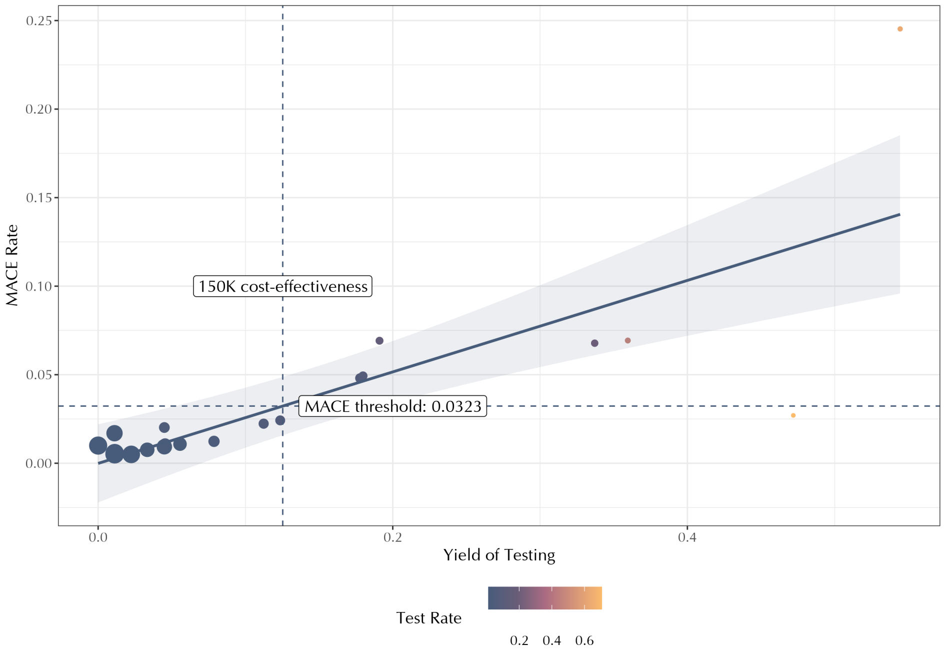
from the 2% adverse outcome rate from heart attack testing guidelines.

3 Counterfactual Testing Policies: Stress Testing vs. Catheterization

In the main text, we treat the two kinds of tests doctors can do—stress tests and catheterization—as one testing decision. We do so for simplicity, and because both types of test are aimed at diagnosing the same underlying coronary blockage. However, the costs and benefits of these two categories of tests are different. This means there is more than one counterfactual we might wish to consider in our analyses of testing decisions. Our main analyses consider a specific counterfactual: eliminating (or adding) the specific kind of test the physician chose to do in our dataset, in a patient with a given level of predicted risk.

Here we present results for other counterfactuals of interest, for example, eliminating all stress tests. Table A.3 shows that, irrespective of risk, the value of stress testing is extremely low. This supports prior literature, which has argued that these tests are sufficiently low value that they should be eliminated completely; we find that there are no groups, even the highest-risk groups, in whom these tests (as physicians currently use them) are cost-effective. By contrast, Table A.4 shows that the value of catheterization is very low in the lowest-risk quintiles, but is cost effective for the highest-risk three quintiles. Finally, while we cannot fully simulate substitutions between testing types—for example, we do not know what stress tests would have shown in catheterized patients—we can calculate a conservative bound on substitution of catheterization for stress tests. Table A.5 shows a counterfactual policy in which we drop only negative stress tests (those that do not progress to catheterization). This simulates a near-perfect strategy of replacing current stress

Figure A.2: Cost-Effectiveness in the Tested vs. Adverse Event Rate in the Untested



tests with catheterization. This strategy has similar cost-effectiveness to catheterization, albeit slightly lower (since a positive stress tests necessarily progresses to catheterization).

Table A.3: Yield, Frequency, and Cost-Effectiveness of Stress Testing

	Yield Rate (SE)	Initial Stress Test Rate (SE)	Cost Effectiveness ($) (Lower-Uppper Bound)
Risk Quintile			
1	0.007	0.011	–
	(0.005)	(0.001)	–
2	0.007	0.014	922,497
	(0.005)	(0.001)	(745,691-1,209,205)
3	0.011	0.035	805,446
	(0.006)	(0.002)	(649,409-1,060,183)
4	0.024	0.051	1,575,410
	(0.011)	(0.003)	(1,205,093-2,274,279)
5	0.053	0.161	465,476
	(0.018)	(0.012)	(374,829-613,953)

Table A.4: Yield, Frequency, and Cost-Effectiveness of Catheterization

	Yield Rate	Initial Stress Test Rate	Cost Effectiveness ($)
	(SE)	(SE)	(Lower-Uppper Bound)
Risk Quintile			
1	0.055	0.002	646,351
	(0.031)	(0)	(505,405-896,313)
2	0.139	0.004	223,991
	(0.039)	(0)	(180,863-294,128)
3	0.238	0.013	132,517
	(0.043)	(0.001)	(108,728-169,632)
4	0.372	0.038	89,606
	(0.039)	(0.003)	(74,010-113,530)
5	0.674	0.24	40,253
	(0.031)	(0.014)	(33,409-50,622)

Table A.5: Yield, Frequency, and Cost-Effectiveness of Substituting Catheterization for Stress Testing

	Yield Rate	Initial Stress Test Rate	Cost Effectiveness ($)
	(SE)	(SE)	(Lower-Uppper Bound)
Risk Quintile			
1	0.056	0.002	885,816
	(0.027)	(0)	(677,765-1,278,171)
2	0.138	0.005	234,344
	(0.036)	(0)	(189,251-307,650)
3	0.212	0.016	155,356
	(0.038)	(0.001)	(127,163-199,612)
4	0.326	0.045	103,017
	(0.035)	(0.003)	(84,975-130,786)
5	0.598	0.275	44,268
	(0.031)	(0.015)	(36,727-55,707)

4 Modeling Approach and Hyperparameters

Most risk prediction tools for heart attack in the medical literature use a handful of clinical variables as predictors, for example, elements of the medical history, certain laboratory studies, or interpreted features of the electrocardiogram (e.g., TIMI, GRACE, or HEART scores). As noted above, claims or electronic health records, by contrast, contain a vast set of other potential predictors that are increasingly being used as inputs to machine learning models in medicine.[4] Building on this work, we design a machine learning algorithm to accurately predict risk out-of-sample using a wealth of EHR data.

4.1 Predictors

To form the inputs to our predictive model, we begin by transforming the discrete person-date data described above (e.g., a cholesterol value of 200mg/dL, or a hospitalization for heart failure, on a given day before the ED visit) into summary statistics (counts, averages, standard deviations, etc.) over discrete time periods (i.e., 0–1 months, 1–12 months, and 12–24 months prior to a visit). For diagnosis and procedure codes, as well as medications, we additionally take advantage of the fact that these codes are nested in categories: we aggregate them into clinically meaningful 'super-variables,' by collapsing at the level of hierarchical taxonomies defined by the Agency for Healthcare Research and Quality's Clinical Classification Software (with minor modifications, available on our online code repository), and the ATC classification for medications.[5] This results in one variable for each time period, describing occurrences over short, medium, and long windows before a given visit, and for each semantically grouped diagnosis, procedure, or medication group. We dropped variables missing in over 99% of the training set, leaving a vector X_{ij} of 16,381 predictors for visit j by patient i.

We were very careful to form these variables so that the information they contain was uniformly available to the physician at the time of the decision. This is harder than it seems: for example, sometimes physician notes are dated on the day of the ED visit, but are completed by the physician days or even weeks later—after information on the results of testing become available. The data available during the course of the ED visit, like the results of laboratory testing or the electrocardiogram, are likewise 'downstream' from the decision making process we aim to assess: only patients suspected of a heart problem will have certain test results present—but we wish to create predictions irrespective of whether the physician suspected a heart problem. So we stop incorporating any information from the EHR starting the moment the patient arrives at the ED triage desk. Later, we will use some data from the ED visit itself to infer the physician's level of suspicion for heart attack, but we do not include any of these data into the predictive model.

[4] Rajkomar, Dean, and Kohane (2019) provide a helpful review of recent work. Notable examples include Ghassemi et al. (2014), Rajkomar et al. (2018), and Henry et al. (2015), who use these tools to predict clinical outcomes, and Miotto et al. (2016) who predict a variety of future diagnoses. We are necessarily brief in our description of machine learning methods; see Mullainathan and Spiess (2017) or Athey and Imbens (2019) for a more thorough overview with references.

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

4.2 Training Procedure

Our goal is to form estimator $\widehat{m}(\cdot)$, on the basis of observed covariates X_{ij}. Our dataset gives us two ways to proxy our (unobserved) quantity of interest, whether or not a patient has a blockage: a positive test result leading to treatment S_{ij} when $T_{ij} = 1$, and an adverse event A_{ij} when $T_{ij} = 0$.

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

In the training set, we form four individual estimators, two in the tested and two in the untested, that will later be combined into an 'ensemble' estimator $\widehat{m}(X)$. First, in the tested patients, we fit two distinct machine learning models, gradient boosted trees and LASSO, both designed to handle large sets of correlated predictors (Friedman, 2001) to predict treatment $S_{ij} = 1$ using observed covariates X_{ij}. This results in two estimators that predict treatment in the tested, one gradient boosted tree and one LASSO.[6] In the untested patients, we fit two similar models to predict $A_{ij} = 1$ using X_{ij}. We do so because, as noted above, we would expect there to be signal for predicting B_{ij} in both treatment S_{ij} when $T_{ij} = 1$ and adverse event A_{ij} when $T_{ij} = 0$. So a model that was useful for predicting one would have signal for predicting the other. This also let us take advantage of the far larger sample size in the untested.

To tune the parameter set for both types of models, we first randomly divide our sample of patients into five folds for cross-validation. The LASSO's optimal value of lambda is determined through cross-validation with R's glmnet function. The loss function we minimize is AUC. The selected lambdas for the yield and adverse event LASSOs are 0.0112 and 0.0053, respectively. The hyperparameters for the gradient boosted trees are selected from large tuning grids. Table A.6 presents the tuning grid for the Yield model, and Table A.7 presents the grid for the adverse event model. We tune with separate grids here because the number of observations in the yield model is an order of magnitude smaller than the number of observations in the MACE model. By cross-validating in the training set, we determined the optimal hyperparameter values (bolded) and fit our gradient boosted models using these values.

Once these models are tuned, we use them to generate predictions (one for each model) in a 5% ensembling set, separate from the training set. We then use logistic regression to ensemble these predictions together, resulting in a calibrated final model of yield in the tested. Table A.8 presents the results from our ensembling regression, which determines the final weights on the LASSO and boosted trees (GBM) predictions to generate our final risk estimates. The output of this weighted combination forms the final ensemble model $\widehat{m}(X)$.

[6] A trivial way to see the benefit of machine learning methods here is that we have only 5,755 tested patients in

Table A.6: Tuning Grid and Selected Parameters: Yield of Testing

Eta	Max Tree Depth	Observation Subsample	Feature Subsample	Min Child Weight
0.05	8	0.75	0.5	20
0.05	8	0.85	0.5	20
0.05	9	0.75	0.5	20
0.05	**9**	**0.85**	**0.5**	**20**

Table A.7: Tuning Grid and Selected Parameters: Adverse Events

Eta	Max Tree Depth	Observation Subsample	Feature Subsample	Min Child Weight
0.05	7	0.5	0.5	20
0.05	7	0.75	0.5	20
0.05	8	0.5	0.5	20
0.05	8	0.75	0.5	20
0.05	**9**	**0.5**	**0.5**	**20**
0.05	9	0.75	0.5	20

Table A.8: Sub-predictor Coefficients in Ensemble

	Yield
Yield GBM	0.0966
	(0.101)
Yield LASSO	1.5846***
	(0.393)
Adverse event GBM	0.334***
	(0.116)
Adverse event LASSO	−0.954***
	(0.286)
Observations	404
Note:	*p<0.1; **p<0.05; ***p<0.01

our training set, and 16,381 predictors, so traditional functional forms are not possible.

5 'Naive' Estimates of Testing Yield

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

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

To illustrate the magnitude of this problem, we first calculate $\tilde{S}$, the yield the model would predict in untested patients at a given level of predicted risk. We find that the untested as a whole have a predicted yield of 4.6%. By comparison, the realized yield in the tested is 14.3%. However, because of the far larger size of the untested set—the entire tested set would make up only 3.0% of the untested set—there are many more patients with very high predicted risk in the untested. For example, the model predicts that 2,797 untested patient-visits would have led to treatment, had they been tested, while in the tested set, physicians actually found only 255 patient-visits that led to treatment. This would imply that doctors are missing 91.6% of all heart attacks among patients passing through the ED. These back-of-the-envelope calculations suggest that model predictions may be missing important signals, and over-predicting risk.

5.1 Capturing Risk Information from Electrocardiograms

To show this more precisely, we turn to a subset of patients in whom we have the opportunity to add an important source of physician private information: the electrocardiogram (ECG). The ECG is a fundamental part of how physicians form their risk estimates on heart attack. Among the 29.4% of patients with ECG results in the ED, the physician can observe an important source of signal regarding the likelihood of blockage. So far, we have not included ECG data in our model, for a specific reason: not all patients have an ECG, and whether they have an ECG is itself a function of the physician's risk prediction. So it would be problematic to rely on these data for a general risk prediction function—it would not be applicable to all patients (i.e., we could not generate predictions for the 70.6% without ECGs). For the purposes of quantifying the physician's informational advantage, however, the ECG is very valuable. This is particularly true because ECG data are rarely included in statistical risk models: even if they were available for all patients in a cohort, they are housed in separate databases from standard EHR data and thus difficult to access. Further, they consist of waveforms that cannot be accommodated easily in traditional statistical risk models. So our results translate more broadly into a range of prediction models in the literature that do not incorporate such important sources of signal available to physicians.

For those patients with ECG data available, we first explore how physicians appear to be using obvious features of the ECG that are suspicious for heart attack, using the physician interpretation entered into the electronic record to accompany the waveform. One caveat to consider is that the formal interpretation we observe is entered not by the emergency physician most directly involved in the diagnostic process in the ER, but rather by a cardiologist who enters the interpretation ex post (for reasons related to hospital billing). So while these two interpretations are correlated, in the sense that physicians all read ECGs in largely similar ways, we do not observe the precise interpretation of the ECG as it is used in the emergency physician's decision process. So we observe how a cardiologist enters her interpretation of the ECG: as a brief, semi-structured free text narrative, for example, "Normal sinus rhythm. Non-specific T-wave abnormalities in the inferior leads."

Using regular expression matching, we turn these interpretations into a vector of features, and specifically focus on six important features of the ECG that arouse suspicion for heart attack. This lets us form a vector of indicators indexing these six findings, as well an indicator for the cardiologist interpreting the ECG as basically normal (i.e., no indication of heart attack or other problems). We then run two parallel regressions exploring the relationship of these seven factors to both the testing decision, and the yield of testing, conditional on predicted risk:

$$T_{ij} = \beta_0 + \widehat{m}(X_{ij})\beta_1 + \mathsf{ECGFeatures}_{ij}\beta_3 + \epsilon_{ij}, \tag{1}$$

$$S_{ij} = \gamma_0 + \widehat{m}(X_{ij})\gamma_1 + \mathsf{ECGFeatures}_{ij}\gamma_3 + \epsilon_{ij}. \tag{2}$$

We find that, of the two features highly suspicious for heart attack, both are highly predictive of testing—and both are additionally highly predictive of yield. The feature of 'ST elevation,' for example, makes testing 3.4 times more likely (7.2 p.p., SE: 2.5), and nearly triples the likelihood that testing will yield a positive result (24.7 p.p., SE: 6.9). Having an entirely normal ECG, by contrast, makes testing 0.9 p.p. less likely (SE: 0.4), and reduces yield of testing by 2.1 p.p. (SE: 1.4). Full results are in Appendix Table A.9.

5.2 Incorporating ECG Waveform Data

While these individual features are clearly meaningful, both for the physician's decision making process and for the likelihood of blockage, we might wish to extract the maximum amount of risk information from the high-dimensional ECG signal, independent of the physician's interpretation. The ECG consists of a 10-second waveform, sampled at 100 Hz. The waveform corresponds to a record of the electrical depolarization of the heart.

To capture this directly, we build a new model of risk, $\widehat{m}_{\mathsf{ECG}}(X)$, that incorporates the ECG waveform, using a convolutional neural net. We use as inputs three waveforms from the ECG, collected at three different points on the chest (leads II, V1, V5). We then feed this into a 34-layer residual neural network, a variant of the standard convolutional neural network used for deep learning, modeled on the architecture in Rajpurkar et al., 2017. The model is trained and applied

Table A.9: Predicting Yield and Test with ECG Features

	Test	Yield
	(1)	(2)
Predicted Risk	1.085***	0.961***
	(0.045)	(0.093)
Normal ECG	−0.009**	−0.021
	(0.004)	(0.014)
Highly Suspicious		
ST Depression	0.080***	0.199***
	(0.030)	(0.075)
ST Elevation	0.072***	0.247***
	(0.025)	(0.069)
Nonspecific Findings		
T-Wave Abnormality	0.041***	−0.017
	(0.007)	(0.018)
T-Wave Inversion	0.033***	0.022
	(0.013)	(0.034)
Bundle Branch Block	0.032***	0.052
	(0.012)	(0.035)
Left Ventricular Hypertrophy	0.008	−0.085***
	(0.010)	(0.023)
Observations	13,609	1,261
R^2	0.098	0.223

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

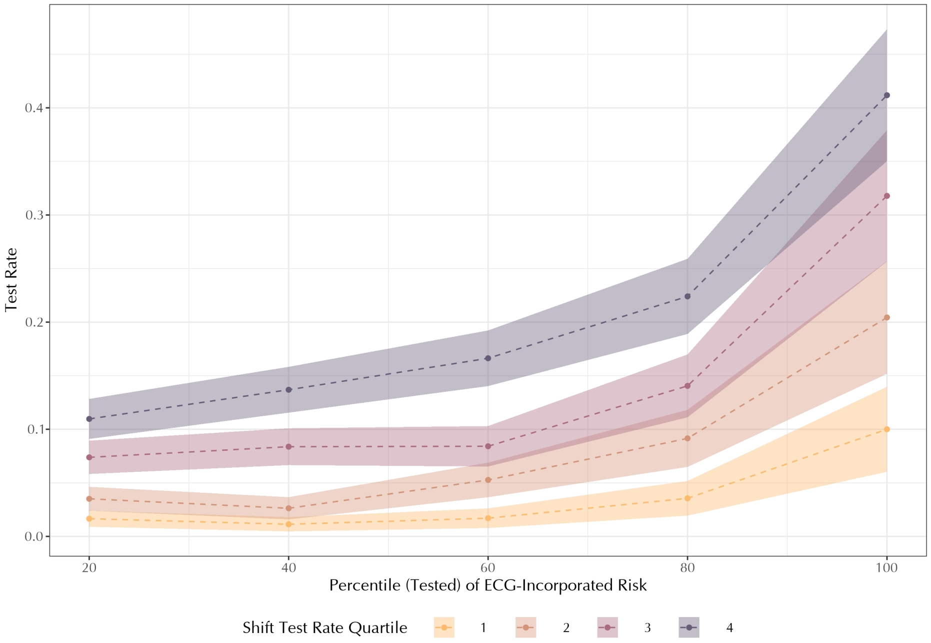
in the same way as our usual risk predictor, $\widehat{m}(X)$, in the sense that it predicts yield of testing in the tested. However, because our tested sample is very small relative to the usual samples needed to train convolutional neural nets, we take two steps to improve model fitting. First, we pre-train the model to predict the vector of cardiologist labels described above, on the full training set. This initializes the model parameters to values that capture meaningful signals. We then create a composite outcome, treatment in the tested set or adverse event in the untested set, that is definable in the entire sample. We train the model to predict this outcome, using as inputs both the risk predictions from our usual model $\widehat{m}(X_{ij})$ (which we form via five-fold cross validation in the training set), and the three-channel raw ECG signal. We include in the final ('fully connected') layer a variable reflecting whether or not the patient is tested, which is incorporated into the final model prediction. As with our usual model, the entire process happens in the training set, and the model is then applied to generate risk predictions in the hold-out.

We then simply compare how much the addition of ECG risk information updates the original prediction, by subtracting $\widehat{m}_{\mathsf{ECG}}(X) - \widehat{m}(X_{ij})$. Across the entire population, we find that adding ECG information decreases risk for 97.6%, and increases it for 2.5%. In the highest-risk bin of untested patients (0.21%) based on $\widehat{m}(X_{ij})$, adding the ECG information decreases predicted risk for 100%, resulting in 24.7% being dropped out of that highest-risk bin. It is revealing that predicted risk *on average* falls. Even if ECGs resulted in better prediction, why is the average prediction across the whole population changing so sizeably? The reason is intimately tied to the physician's information advantage. We are training models on the tested and (naively) extrapolating predictions to both tested and untested. If the physician is using unobservables to select who is tested, then the untested are less risky—even conditional on observables—than the tested. So a better predictor that uses these unobservables, will predict lower overall risk, which we see here.[7]

5.3 Physicians' Marginal Patients and Risk Information from the ECG

Given the substantial amount of private information available to physicians in the ECG waveform, an important question is whether our analysis of the risk of physicians' marginal patients is robust to the inclusion of this information. In other words, do physicians still draw their marginal patients from across the risk distribution, even taking ECG information into account? Figure A.3 shows that this is still the case. An important caveat is that we can only produce these predictions for the subset of patients with ECG information available.

Figure A.3: Variation in Testing Rates by Predicted Risk, Incorporating ECG Waveform



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

6 Adverse Outcome Robustness Checks

6.1 Troponin Thresholds

To ensure that diagnosed adverse events like heart attack are not discretionary diagnoses or incentivized ‘up-coding,’ we use the laboratory data present in electronic health records to confirm heart attack, using a test called troponin. Troponin is a protein found in heart muscle that is released into the blood stream if heart cells are damaged, where it is detectable and can indicate heart attack. However, elevated troponin can also result from other causes. One may worry that troponin elevations, particular at lower levels, might be the result of these other causes (e.g., patients with renal failure often have detectable troponin rates despite not having a heart attack). While we believe this is unlikely, since our definition of adverse event also requires the physician to have documented heart attack, in Figure A.4, we show adverse rates for progressively more stringent thresholds ($Tn \geq 0.05$, $Tn \geq 0.1$, and $Tn \geq 0.5$). Particularly at higher levels, troponin elevations are more likely to represent true—and severe—heart attack. We find that, even if we restrict to what would be a large amount of damage to heart muscle and very suspicious for heart attack, high-risk patients still have diagnosed adverse events at high rates.

We also use the results of the troponin test to exclude patients in whom physicians may suspect heart attack, on the basis of a positive test, even if they do not document heart attack in the record. To provide some sense of the severity implied by different levels of troponin, and the ways in which physicians use troponin to guide their testing decisions, Figure A.5 provides rates of key outcomes (testing, yield, and AMI diagnosis) as a function of the troponin levels detected on the day of the emergency visit (using the maximum, if more than one value was found).

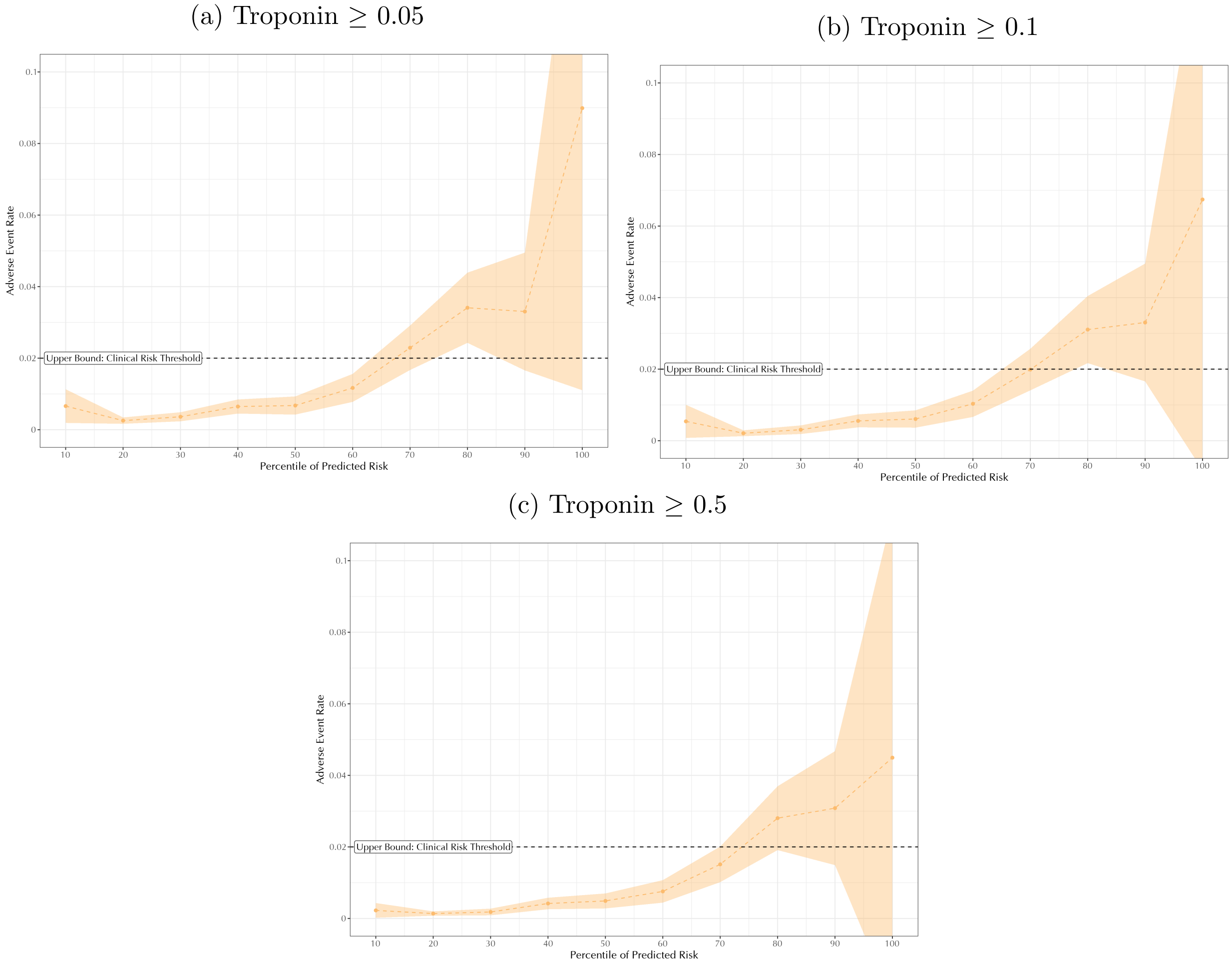
6.2 Excluding Same-day ECGs and Troponin Tests

As described in the main text, we may exclude patients with ECGs or troponin tests in the ER to get a lower bound on the number of untested patients in whom physicians never suspected heart attack. Looking at 30-day adverse outcomes in this restricted sample, we still find evidence of a high rates of adverse events, well above the clinical threshold of 2%, for patients in the highest-risk bins. We offer tabular results to supplement Figure 3 in the main text, both for the sample excluding patients with ECGS (column (3)) and the sample excluding patients with troponin tests in the ED (column (4)). These results are in Table A.10. As in the main text, the risk bins are formed from quintiles of predicted risk in the tested for comparison.

6.3 Sensitivity Analysis: Patients with Alternative Diagnoses (Not Heart Attack)

Figure A.6 is an analogue to Figure 2 in the main text, but here we exclude patients who were admitted to the hospital with uncertain diagnoses. The intuition here is that physicians might suspect blockage—just not enough to document it explicitly. In this case, the physician has not reached or documented another diagnosis (like pneumonia), but rather a symptom-based diagnosis (like chest pain) that indicates she has not reached a conclusion on what is going on with the patient. Of course, not all these patients will go on to have heart attack, but by excluding them we are left with a population in whom the adverse event rate is unlikely to reflect private information. The physician has explicitly reached the conclusion that something besides heart attack is going on. So in these remaining patients, it seems unlikely that the physician has suspected them strongly of

Figure A.4: Adverse Outcome Rates For Various Troponin Thresholds


(a) Troponin ≥ 0.05

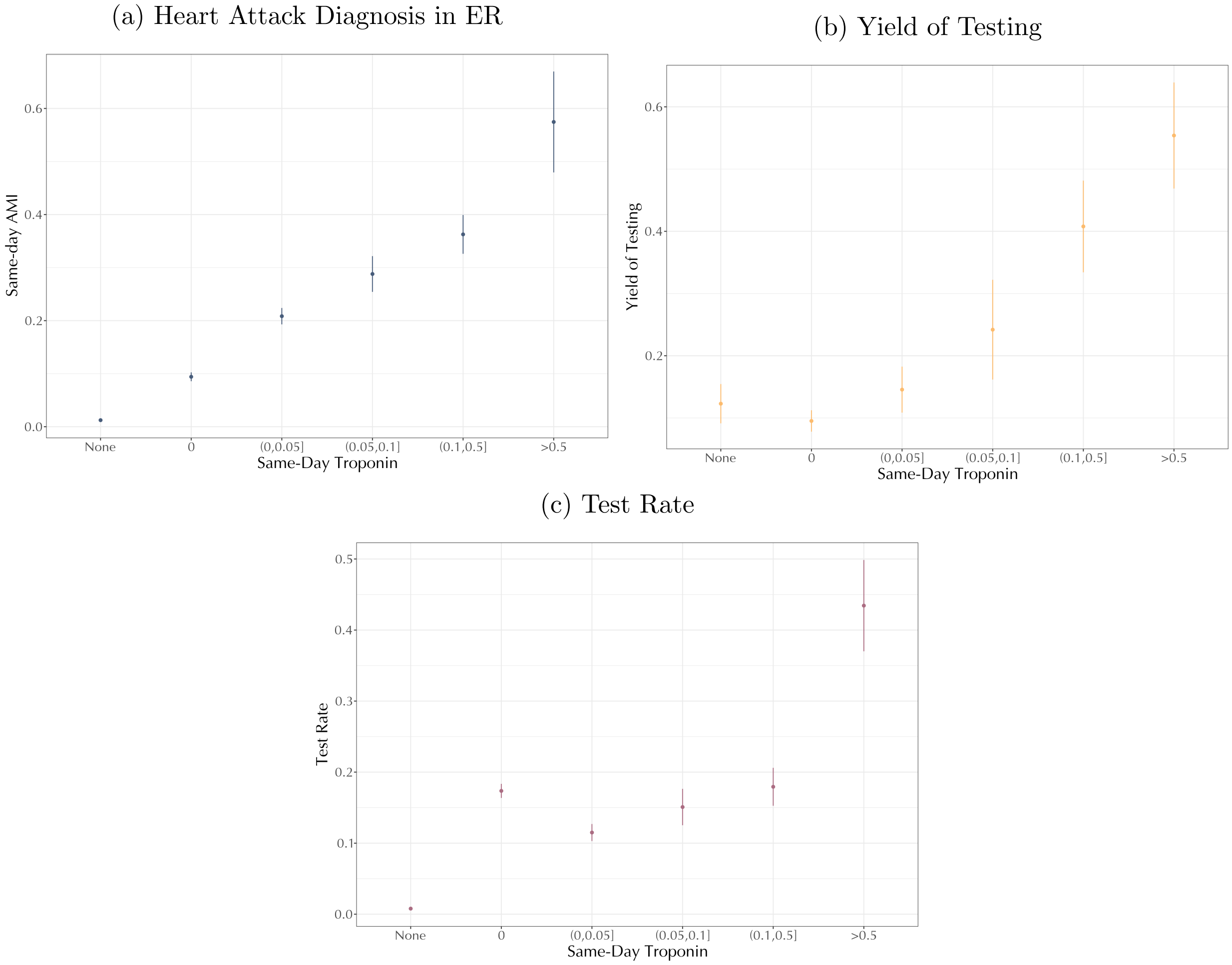
(b) Troponin ≥ 0.1

(c) Troponin ≥ 0.5

heart attack, and also concluded that they are unsuitable for invasive testing or treatment. We also include in this sample those patients who were sent home from the ER, in whom the physician felt there was such a low risk of anything serious that they were not worth admitting to the hospital for observation (again, not something physicians do when they suspect heart attack).

Despite this, we see very similar adverse event rates in this population to the main population: a 5.66% rate of adverse events in the highest risk bin. This, in combination with our results in patients who did not have an ECG done in the emergency department, is reassuring: the physician is unlikely to have suspected heart attack, and thus even less likely to have ruled out treatment because the patient could not tolerate it.

Figure A.5: Outcomes for Patients Based on ED Troponin Tests



(a) Heart Attack Diagnosis in ER

(b) Yield of Testing

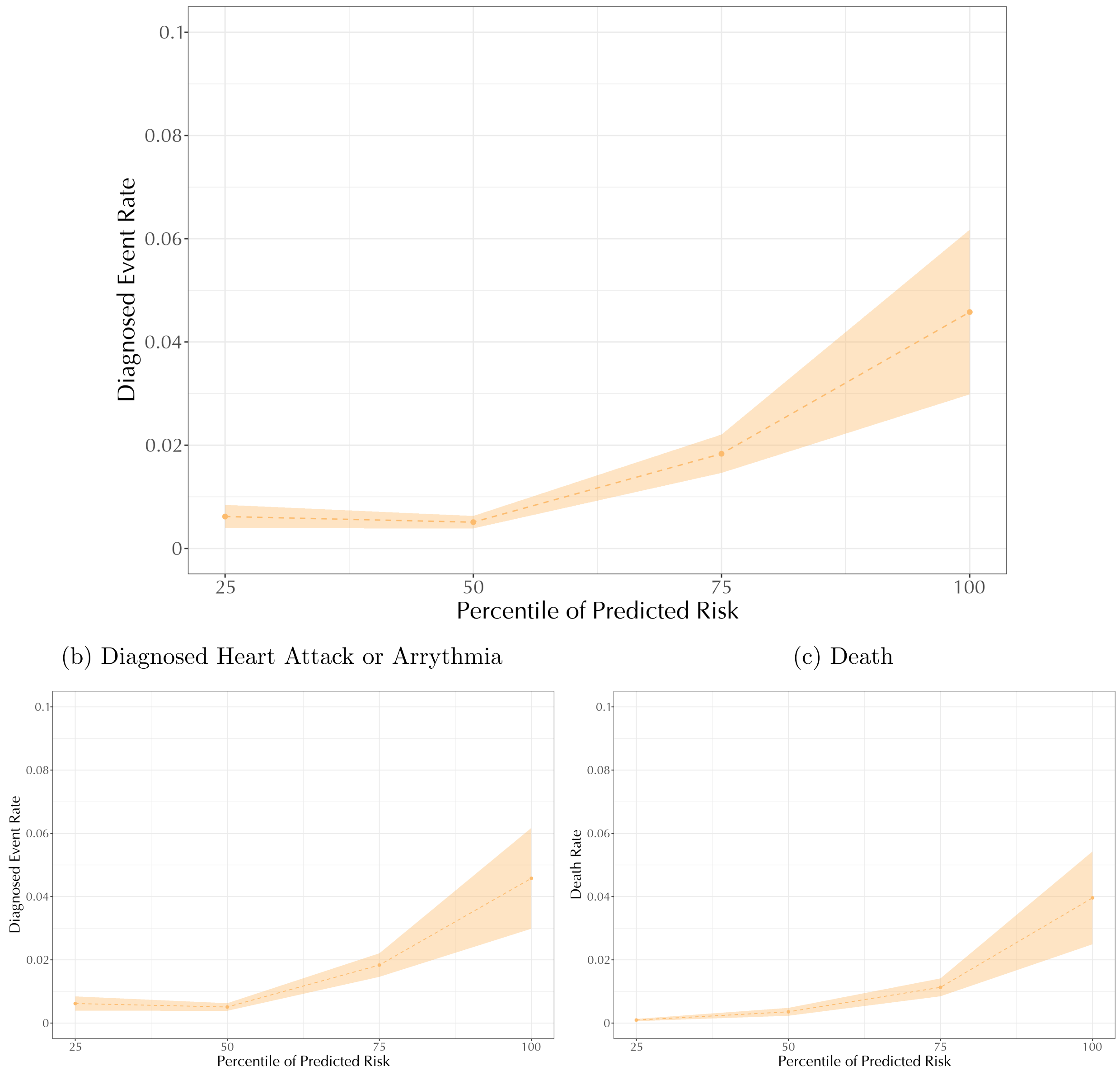
(c) Test Rate

Table A.10: Rates of Adverse Cardiac Events and Death

.

			Adverse Event Rate	Adverse Event Rate	Adverse Event Rate
		Mean Risk	(All)	(No ECG)	(No Troponin)
		(1)	(2)	(3)	(4)
Risk Bin					
1	0.019	0.008	0.004	0.007	
	(0)	(0.001)	(0.001)	(0.001)	
2	0.046	0.006	0.002	0.005	
	(0)	(0.001)	(0)	(0.001)	
3	0.084	0.011	0.004	0.009	
	(0)	(0.001)	(0.001)	(0.001)	
4	0.148	0.027	0.016	0.021	
	(0)	(0.003)	(0.003)	(0.003)	
5	0.316	0.059	0.042	0.063	
	(0.014)	(0.013)	(0.019)	(0.017)	

Figure A.6: Adverse Outcomes in Patients with Other Diagnoses Besides Heart Attack



(a) Any Adverse Event

(b) Diagnosed Heart Attack or Arrythmia

(c) Death

7 Natural Experiment: Additional Checks

7.1 Exclusion Restriction

In the main text, we claim that shift triage teams are a valid instrument for physician testing decisions. One may be concerned that testing "quotas" could violate the exclusion restriction. To test for this, we construct three new variables for each encounter: the number of tests in the previous 12, 24, and 48 hours respectively. Then, we regress testing decision on risk, as well as these variables. We find that recent testing volume is unrelated to a physician's decision to test.

Table A.11

	Test		
	(1)	(2)	(3)
Predicted Risk	0.546***	0.546***	0.546***
	(0.014)	(0.014)	(0.014)
Tests in Previous 12 Hrs	0.0002		
	(0.0002)		
Tests in Previous 24 Hrs		0.0001	
		(0.0001)	
Tests in Previous 48 Hrs			−0.00002
			(0.0001)
Observations	234,134	234,134	234,134
R^2	0.046	0.046	0.046
Note:	*p<0.1; **p<0.05; ***p<0.01		

7.2 Test Prediction and Balance on Yield

Appendix Table A.12 shows the results of four regressions. In Column (1), we regress the outcome of testing on the shift's (leave-one-out) testing rate ($\bar{T}_{-j}$), controlling for time fixed effects (year, week of year, day of week, and hour of day) and patient risk. In Column (2), we add an interaction term between the shift's testing rate and predicted yield. These regressions test whether patients in high-testing shifts are riskier on unobservables—in particular, if providers pick up on this unobservably higher risk and test more as a result, patients in higher-testing shifts should have higher yield than expected based on risk. In fact, there is no significant correlation between testing rate and yield, either alone or in the interaction. While estimates are imprecise, they do argue against large imbalance on unobservables.

In Column (3), we regress an individual's testing dummy (T_{ij}) on the shift's (leave-one-out) testing rate ($\bar{T}_{-j}$), controlling for time fixed effects (year, week of year, day of week, and hour of day) and patient risk. We find that a one-standard-deviation increase in shift testing rate (2.3 percentage points) increases individual testing probability by 0.19 percentage points (SE: 0.06), or 6.7% of the base test rate. In Column (4), we add an interaction between the shift's testing rate and predicted yield. We find a positive interaction term, implying that higher-testing shifts test higher-risk patients significantly more than lower-risk patients. This mirrors the results in Figure 7.

Table A.12: Shift Test Rate and Yield, Test

	Yield		Test	
	(1)	(2)	(3)	(4)
Predicted Risk	0.923^{***}	0.912^{***}	0.564^{***}	0.515^{***}
	(0.039)	(0.061)	(0.014)	(0.023)
Shift Test Rate	0.308	0.248	0.082^{***}	-0.043
	(0.222)	(0.282)	(0.028)	(0.042)
Predicted Risk × Shift Test Rate		0.335		1.604^{***}
		(1.488)		(0.618)
Observations	4,241	4,241	123,289	123,289
R^2	0.176	0.176	0.058	0.058
Note:		*p<0.1; **p<0.05; ***p<0.01		

7.3 Random Effects Model of Shift Testing Rates

We calculate shift-level testing propensities via the following model:

$$T_{ij} = \beta_0 + \mathsf{Shift}_j\beta_1 + \widehat{m}(X_{ij})\beta_2 + \mathsf{TimeControls}_j\beta_3 + \epsilon_{ij}. \quad (3)$$

Each visit's testing likelihood is modeled as a function of a vector of indicators indexing the particular triage shift she showed up in, modeled as a random effect. In addition, we control for a vector of time variables for visit j, that captures differences in testing rate attributable to the mix of patients showing up at a given time (i.e., fixed effects for year, week of year, day of week, and hour of day), as well as patients' predicted risk.[8] We end up with $3,951$ random effects that measure the testing propensity of each shift, and verify that the variance of these effects is non-zero (p = 0.0003) by running 1,000 bootstrap simulations.

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

8 Replication of Results in National Medicare Claims

To ensure that our results are not specific to the single hospital we study, we turn to a nationally representative 20% sample of Medicare claims data. We identified 20,059,154 emergency department (ED) visits over a four-and-a-half-year period from January 2009 through June 2013 (we use the last half year of 2013 as a follow-up period for included visits). We excluded non-fee-for-service patients, since we do not observe their full claims history.

8.1 Sample

We apply similar exclusions to the same as we describe in the electronic health record sample (except as noted in the main text): those who died in the ED, visits preceded by recent known heart attack or its treatment, those whose general poor health might limit the benefit of testing or treatment (80 years of age or older, poor-prognosis conditions, hospice or nursing home care, etc.). Our final sample contains 4,246,642 visits: 189,290 visits in which patients were tested, and 4,057,352 in which they were not. Of the tested, 124,736 had stress tests, and 84,481 had cardiac catheterization; 13,930 had both a stress test and subsequent catheterization. Among the tested, we identified 24,126 who received stents. Summary statistics on demographics are shown in Table A.13. As usual, we exclude untested patients who were diagnosed with conditions related to heart attack on the day of or the day after their ER visit.

Table A.13: Medicare—Sample descriptive statistics.

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

8.2 Main Results

Figure A.7 shows the realized yield of testing against bins of predicted risk, and the cost-effectiveness of tests by risk bin. At a threshold of $150,000 per life year, 52.6% of tests can be flagged as wasteful. Again, individuated risk predictions are key here: had we taken the typical approach of looking only at average yield, we would have concluded that testing overall was somewhat cost-effective, at $135,859 per life year. Panel (a) of Figure A.8 shows the rate of diagnosed adverse events and death in the 30 days after visits, which increase in predicted risk. The highest-risk bins are well above the 2% threshold: 3.8% have diagnosed heart attack or arrythmia, and an additional 1.5% die.

To translate this into an estimate of under-testing in the absence of a quasi-experiment like shift-to-shift variation, we estimate a simple lower bound. Observe that the lowest rates of under-testing would be if all adverse outcomes were concentrated (with $p = 1$) in an ex ante identifiable set of people. In other words, a conservative estimate of under-testing is to consider only those untested high-risk patients who go on to have realized adverse events, as those who would have been likely to benefit from testing. To estimate the cost effectiveness of testing them, we simply use the cost effectiveness implied by their ex ante risk (estimated in the tested). We can then calculate, at different thresholds for cost effectiveness, the net amount of over- and under-testing at a cost per life year valuation of $150,000. As noted above, we would drop the 52.6% of tests doctors currently do, but we would also add back 17.9% (relative to the current number of tests) for high-risk patients not currently tested. Importantly, even though this strategy would on net *reduce* testing, a large fraction of the benefits of this reallocation comes from *increasing* testing for the high-risk untested. For example, at $150,000 per life year, we would reduce testing by 34.7% on net—but 42.8% of welfare gains come from remedying under-testing (i.e., $228.0 million in surplus from life years saved), as opposed to reducing over-testing (i.e., $304.7 million saved from dropping low-value tests). This fraction of benefits from increasing testing grows with the valuation of a life year.

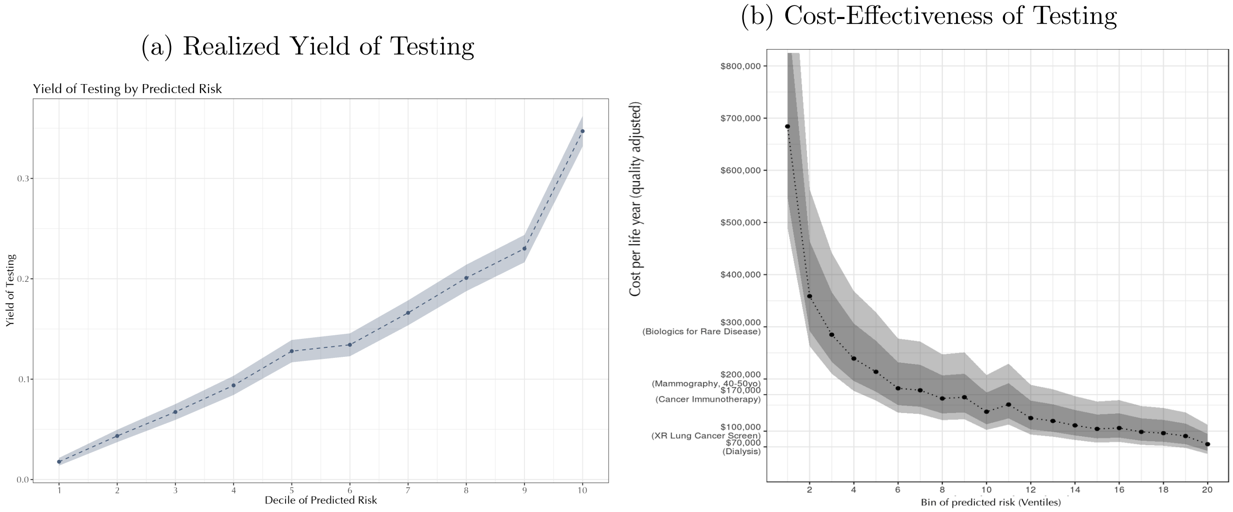
8.3 'Natural Experiment': Weekday vs. Weekend Testing

In the main text, we use a natural experiment exploiting differences in shift-level testing propensities to show that marginal tests come from the entire risk distribution. We replicate this finding in our Medicare claims data, taking advantage of a 'natural experiment' in testing for heart attack that takes place in hospitals across the US every weekend.

Consider the following two facts regarding common practices in cardiac testing. First, it is expensive to maintain staffing of cardiac testing facilities, leading many hospitals to leave them unstaffed on weekends (Krasuski et al., 1999). While testing is still available, if the doctor on duty makes the decision to call in the team from home, it is widely assumed to require a higher threshold for doing so. Second, patients who present to the ED with concerning symptoms on day t are tested immediately to rule out obvious heart attack. If obvious problems are excluded, the decision is made to proceed with cardiac stress testing or catheterization; but this is typically not done on the same day. This is both since it takes time to arrange the test, and because of the need to observe patients for stability: there is an elevated risk of sudden death if tests are done on unstable patients, so patients are typically monitored overnight and undergo repeat laboratory testing (troponin), and have stress tests or catheterizations on day $t + 1$ (or later).

As a result of these two facts, we hypothesized that patients who come in on the day before

Figure A.7: Medicare—Yield and Cost-Effectiveness of Testing in Tested Patients)


(a) Realized Yield of Testing

(b) Cost-Effectiveness of Testing

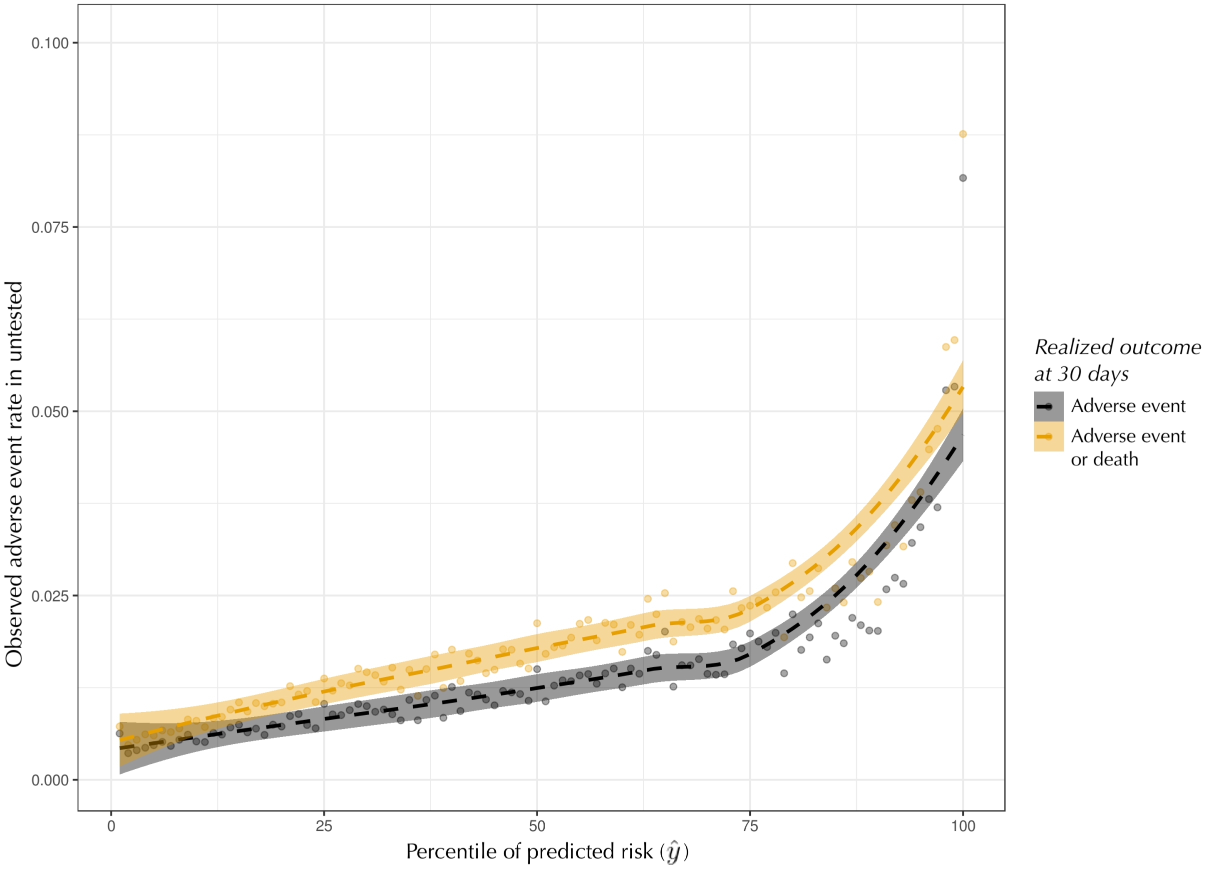
Notes: Realized yield of testing (a) and cost-effectiveness (b) of tests (y-axis) in the tested, by bin of predicted risk (x-axis). Bins are deciles and ventiles of predicted risk. The cost-effectiveness line shows our preferred specification, and the shaded interval shows sensitivity to a range of estimated treatment effects from the literature.

a weekend day—i.e., Friday or Saturday—would be less likely to be tested. This strategy builds on prior research showing differences in care for patients admitted on weekends vs weekdays (Bell and Redelmeier, 2001), but has the additional advantage of straddling a weekend (Saturday) and weekday (Friday), which reduces the risk of confounding by simply comparing weekend patients to weekday patients. To reduce other sources of bias, we also wished to exclude patients who had been transferred or referred to specialized hospitals from other facilities, for whom decision making might be less sensitive to inconveniences related to in-hospital staffing. Practically, we restricted our sample to hospitals with a catheterization laboratory on-site (using the American Hospital Association annual survey data), and to patients whose home zip codes are within 10 miles of these facilities, to zero in on patients presenting to hospitals near their home zip code that had on-site testing facilities. In this sample, we find that patients are 18.3% less likely to be tested when their index visit falls on Fridays and Saturdays than on Sundays through Thursdays (3.80% vs. 4.65%, $p < 0.001$). Panel (a) of Figure A.10 shows that, conditional on geography (i.e., hospital referral region) and year, these patients appear otherwise quite similar on observables. There are small differences in some risk factors for heart disease: while some of these are statistically significant after Bonferroni adjustment, they are substantively small, most on the order of < 0.01 SD units and statistically insignificant. Finally, as a summary statistic, there is only a very small (0.01 SD) difference in overall risk, measured by $\widehat{m}(X_{ij})$, that is also statistically insignificant, meaning that many small differences in individual variables largely balance out.

We first verify that the model predicts accurately in this setting: rates of realized yield in the tested and adverse events in the untested are similar in both pre-weekend and pre-weekday patients. We also verify the relationship between yield (among the tested) and adverse event rates (among

Figure A.8: Medicare—Adverse Events in Untested Patients

Figure A.9: Adverse Events (30 Days After Visits)

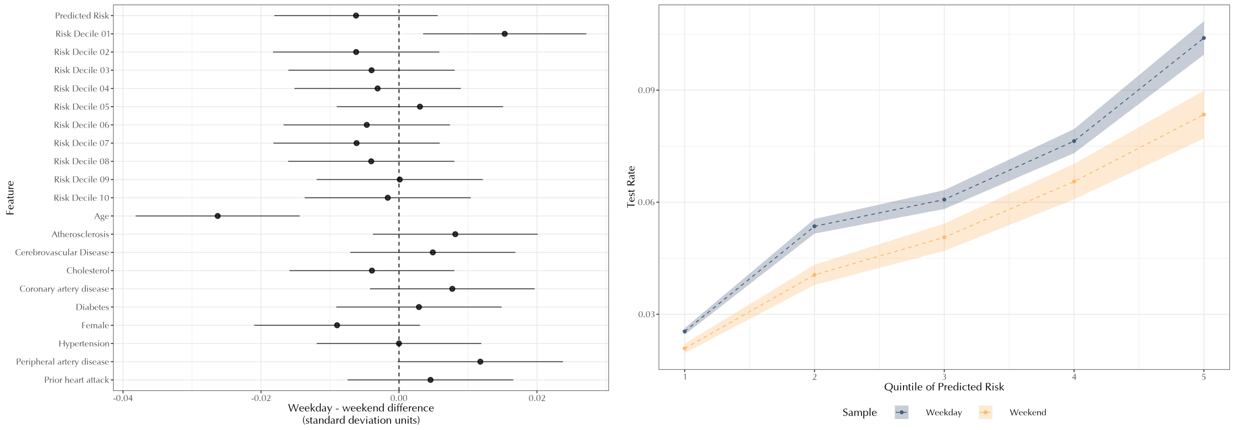


Notes: (a) Rate of adverse events and death over the 30 days following visits (y-axis) among untested patients, by bin of predicted risk (x-axis). (b) Rates of adverse events in the year after visits (y-axis), excluding the first 30 days, for tested (yellow) vs. untested (gray) patients, by bin of predicted risk (x-axis).

the untested), in each percentile bin of $\widehat{m}(X_{ij})$: this is monotonic and approximately linear in this weekend vs. weekday population. This gives us some suggestive evidence that we are measuring the same underlying latent risk, manifested differently depending on whether doctors decide to test or not.

Finally, in this setting with limited influence of unobservables, we can more precisely answer the question: when doctors reduce testing by 18.3%, where in the risk distribution do the marginal patients come from? The results in Panel (b) of Figure A.10 echo the results from our EHR triage shift analysis. We see that when doctors reduce testing on weekends, they drop marginal patients from across the risk distribution, not just low-risk patients. For example, patients in the lowest-risk quintile are 18.1% less likely to be tested on a weekend (2.08% vs 2.54%, $p < 0.001$); patients in the highest-risk decile are 19.7% less likely to be tested (8.34% vs 10.40%, $p < 0.001$). This suggests that, when doctors cut back on testing, they do so fairly indiscriminately.

Figure A.10: Medicare—Weekend vs. Weekday Testing



(a) Balance on Observables, Weekend vs. Weekday

(b) Weekend vs. Weekday Test Rate by Predicted Risk

Notes: (a) Balance checks for a 'natural experiment,' in which patients are tested at higher or lower rates (conditional on geography and year), based on the day of the week they arrive. (b) Testing rates for weekday vs. weekend ED patients. When doctors decrease testing on the weekend, they drop tests across the entire risk distribution.

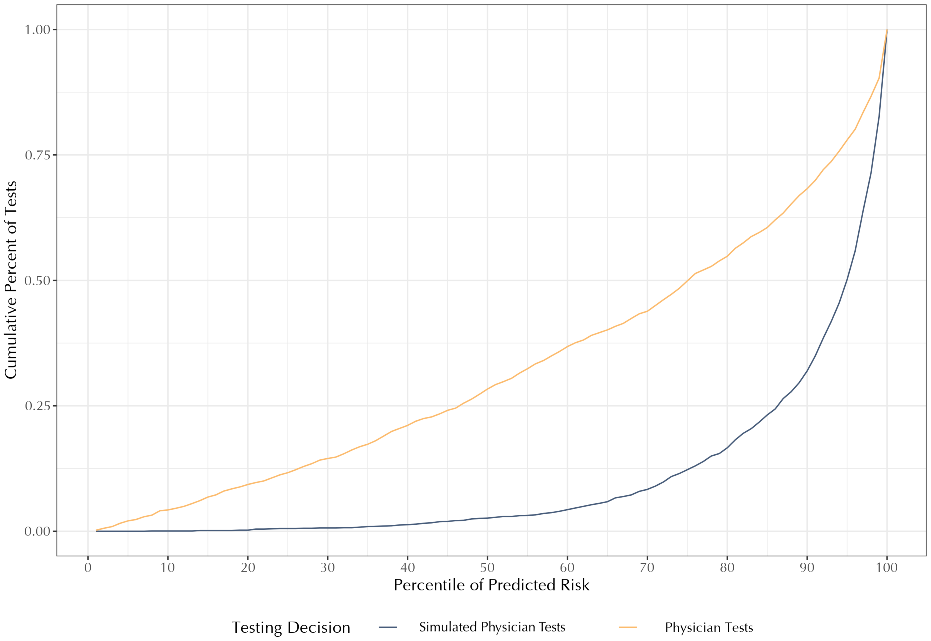
9 Supplemental Analyses of Physician Behavior

9.1 Comparing Algorithm and Physician Testing Decisions

Many of our analyses take the model's predicted risk as a stable object against which physicians' predictions can be compared, whether to quantify the welfare losses from errors, or to establish facts about physician behavior. One potential worry with this setup is that any discrepancies we see between the physician and the model could simply be the consequence of comparing two noisy signals to each other: even if these predictors were equally 'good' (e.g., the probability of blockage conditional on the signal is the same, and the testing rule conditional on the signal is the same), they will differ due to noise.[9] At its extreme, two statistical models estimated on different data would not agree perfectly with each other—there is just variation due to sampling.

To test whether this kind of effect alone could drive our results, we randomly split our train set in half at the patient level. We then train two separate models, one on each half of the data. Model training is otherwise similar to the process used to form our usual $\widehat{m}(X)$: a LASSO and then a gradient boosted tree to predict treatment in the tested, and then an ensemble in a separate set to generate our final scores. Each of these models is used to generate two predictions in the holdout set. This effectively gives us two noisily correlated predictors, of which we choose one (arbitrarily) to play the role of the algorithm, $\widehat{m}_{\mathsf{alg}}(X_{ij})$, and the other to play the role of the physician, $\widehat{m}_{\mathsf{md}}(X_{ij})$.

Figure A.11: Comparing test rates by risk for physician and algorithm testing decisions.



Then, in the hold-out set, we simulate algorithm $\widehat{m}_{\mathsf{md}}(X_{ij})$'s 'testing decisions.' Using the total

[9] It is worth noting that, even if this were the case, there would still be welfare gains from incorporating a second noisy signal (e.g., algorithmic predictions) into decisions that are currently based on only one noisy signal (e.g., physicians' risk predictors). In other words, if physicians are indeed noisy predictors of ground truth, when there is a better (less noisy) predictor of ground truth, the physician would not be optimal.

number of physician tests as a benchmark, we take the 1,834 patients with the highest risk scores according to $\widehat{m}_{\mathsf{md}}(X_{ij})$, and assign them to 'simulated physician testing.' Of course, we also know which patients actual physicians assign to testing. So in Figure A.11, we compare the simulated physician testing to actual physician testing, both with respect to the predictions of our simulated algorithm, $\widehat{m}_{\mathsf{alg}}(X_{ij})$. Exactly as we do in our usual analyses, we bin patients in the hold-out by risk according to $\widehat{m}_{\mathsf{alg}}(X_{ij})$: this is shown on the x-axis. We then show on the y-axis the cumulative fraction of simulated tests and actual tests, by bin of predicted risk. We see that the simulated physician appears to be far better than actual physicians at discerning high-risk patients from low-risk. This indicates that our results do not simply reflect inevitable disagreement between two predictors driven by noise: physicians appear to be leaving far more signal on the table than a simulated physician, even when judged by a predictor with some inevitable disagreements driven by noise.

9.2 Physician Boundedness

In the main text, we provide evidence that physicians use a simplified risk model to make decisions about whom to test. In Figure 5 in the main text, we show how well LASSO models of varying complexity are able to predict test and yield, using R-squared as our performance metric of choice. Here, we provide results from several variations.

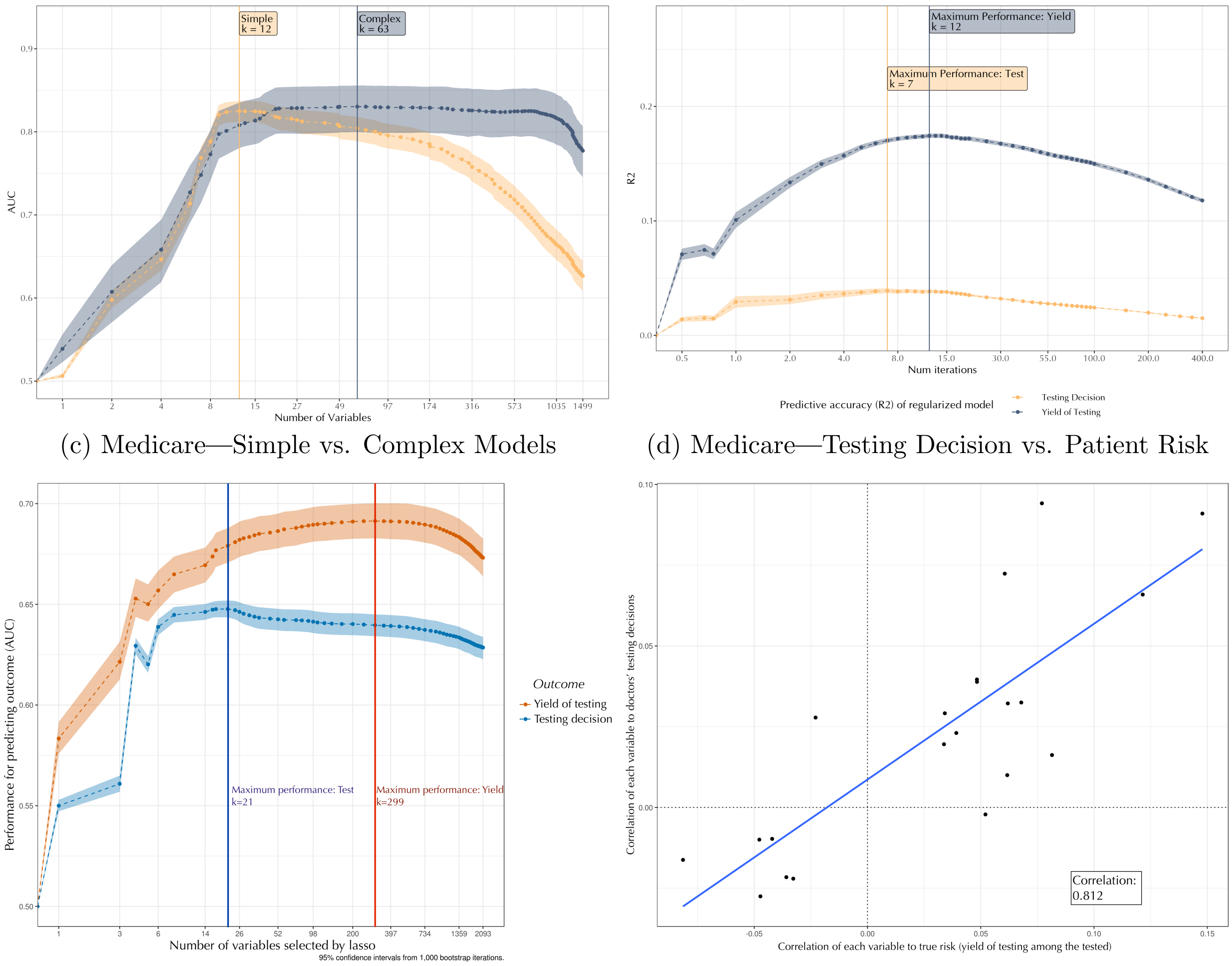
9.2.1 Robustness Checks

In Figure A.12, Panel (a) shows that our results are not specific to the particular metric we use to explain variance: using a measure of discrimination, area under the curve (AUC) as our performance metric, yields similar results. Panel (b) then shows that are results are not specific to a LASSO or linear model: we replicate a similar result using a different model, this time the gradient boosted model that also forms part of our full ensemble model. Instead of number of variables as our measure of complexity, the x-axis now shows the number of boosting iterations of our model (each iteration fits a new model, to the residual from the previous iteration; the iterations are ensembled together at the end, so each iteration adds quite a bit of complexity). The trees used for this model are set to a maximum depth of 5, similar to that used in our main model. This exercise has similar results to the LASSO in terms of complexity: the best model to predict testing decisions is less complex (k_h^*: 7 iterations) than the best model to predict yield of testing (k_r^*: 12 iterations), indicating that physicians use a simplified risk model to make decisions even when the regularization is accomplished with completely different machine learning infrastructure.

9.2.2 Replication: National Medicare Claims

We also replicate this exercise in our Medicare claims data. Panel (c) of Figure A.12 shows LASSO performance (using AUC) for predicting the testing decision and patient risk (yield of testing), based on model complexity. We find that $k_r^* = 299$ variables, while $k_h^* = 21$. Panel (d) reproduces the result from Figure 7 in the main text, showing that although physicians may use a limited set of variables in their risk model, they get the weights on these variables approximately correct.

Figure A.12: Physician Boundedness: Additional Specifications


(a) Simple vs. Complex Models: AUC

(b) Simple vs. Complex Models: Boosted Trees

(c) Medicare—Simple vs. Complex Models

(d) Medicare—Testing Decision vs. Patient Risk

9.3 Physician Biases

9.3.1 Patient Demographics

As we saw in Table 6 of the main text, physicians tend to overweight risk information from patient demographics. In Table A.14 below, we break this out into individual components of patient demographic information: age, race (as reported by the patient), and gender. We find small but significant relationships of certain demographics with testing: older patients and women appear to be tested more than their risk, while self-reported Hispanic patients are under-tested.

9.3.2 Specific Symptom Salience

In Figure 6 of the main text, we saw that the risk weight physicians place on a symptom does not always match the true risk weight. In particular, we noted that physicians generally overweight

Table A.14: Testing and Patient Demographics

	Test	Yield
	(1)	(2)
Predicted Risk	0.809^{***}	0.971^{***}
	(0.060)	(0.090)
Age	0.001^{***}	0.002^{**}
	(0.0001)	(0.001)
Race: Black	-0.003^{*}	-0.032
	(0.002)	(0.022)
Race: Hispanic	-0.004^{**}	-0.013
	(0.002)	(0.024)
Female	0.008^{***}	-0.003
	(0.002)	(0.018)
Low Income	0.012	0.009
	(0.008)	(0.101)
Constant	-0.050^{***}	-0.079^{*}
	(0.004)	(0.046)
Observations	61,135	1,765

Note: *p<0.1; **p<0.05; ***p<0.01

"chest pain" as a symptom when making testing decisions. In Table A.15, we use the ten most common symptoms in our data to predict physician test decisions and yield, controlling for risk.

9.3.3 Representativeness

In the main text, we show that physicians tend to overweight risk signal from symptoms representative of heart attack. To determine which symptoms are 'representative,' we calculate both the frequency of symptoms in those who are found to have heart attacks and the frequency in the entire population. We calculate the representativeness, as the former rate divided by the latter. We consider any symptoms with a ratio greater than 1 (i.e., they are more frequent in those with heart attack than in the overall population) as representative. In Table A.16, we show the representativeness ratios for all symptoms that occur in the set of patients with an eventual yield.

9.3.4 Risk Scores Based on Subsets of Inputs: Correlation with Yield of Testing

In the main text, we show that physician testing decisions are driven in part by a few salient variables. In particular, we saw that physicians tended to overweight risk signal from demographic information, as well as symptoms, especially those representative of patients with heart attacks. In Table A.18, we offer an analogue to Table 8 in the main text with yield as our outcome of interest rather than test. As expected, we see that none of these subcategories of risk are significant in predicting true yield on top of full measure of predicted risk.

9.3.5 Testing Errors: Role of Boundedness and Bias

We measure how much riskier (or less risky) a patient appears if only simple risk is accounted for, by calculating $\widehat{m}_{\mathsf{simple}}(X_{ij})) - (\widehat{m}_{(}X_{ij})$. We then look at the distribution of this variable for both low-risk tested patients (the 'over-tested') and high-risk untested patients (the 'under-tested'). A full 35.5% of the over-tested come from the top quintile, meaning their simple risk is much larger than their actual risk (compared to 14.5% in the lowest quintile). Likewise, among the undertested, 74.2% come from the bottom quintile, meaning their simple risk is much smaller than their actual risk (compared to 7.4% in the top quintile).

Table A.15: Predicting Test and Yield with Individual Symptoms

	Test	Yield
	(1)	(2)
Predicted Risk	0.653***	1.010***
	(0.045)	(0.085)
Abdominal Pain	−0.005***	−0.001
	(0.001)	(0.031)
Chest Pain	0.160***	0.028*
	(0.008)	(0.016)
Shortness of Breath	0.074***	0.040*
	(0.007)	(0.021)
Leg Pain	−0.002	−0.075***
	(0.004)	(0.017)
Foot Pain	−0.007***	0.037
	(0.002)	(0.089)
Back Pain	−0.010***	0.143
	(0.001)	(0.129)
Wrist/Hand Pain	−0.018***	0.457
	(0.001)	(0.363)
Headache	−0.005*	0.057
	(0.002)	(0.078)
Bleeding	−0.005**	0.054
	(0.002)	(0.109)
Car Accident	−0.015***	0.070
	(0.002)	(0.162)
Constant	−0.014***	−0.022*
	(0.002)	(0.013)
Observations	61,965	1,784
R^2	0.150	0.211

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.16: Symptom Frequency and Representativeness for Heart Attack

	Frequency (%)	Representativeness
	(1)	(2)
Representative Symptoms		
Referral: Suspected Heart Attack	0.04	184.80
Arrest	0.05	97.13
Referral for ECG	0.06	39.23
Chest Pain	6.93	9.33
Jaw Complaint	0.17	4.27
Chest Complaint (not pain)	0.73	2.82
Side Weakness	0.22	2.55
Shortness of Breath	3.44	2.18
High Blood Pressure	0.72	1.68
Non-representative Symptoms		
Tachycardia	0.85	0.88
Arm Pain	1.53	0.86
Loss of Consciousness	1.45	0.84
Weakness	1.19	0.79
Nausea	1.67	0.34
Dizziness	2.35	0.32
Abdominal Pain	11.80	0.13

We undertake a similar analysis for patients whose risk comes disproportionately from representative symptoms, i.e., large $\widehat{m}_{\mathsf{represent}}(X_{ij})) - (\widehat{m}(X_{ij})$. An important caveat here is that the representative risk is built only on nine indicator variables and thus does not have a wide range; as a result, we are limited in this analysis. Nonetheless, results are shown in Figure A.14. Those in

Table A.17: Symptoms Recorded at Triage for All 1,069 Treated Patients

Symptom	Count	Symptom	Count
Chest pain	691	Hip, femur, or groin complaint	2
Referral: Suspected ACS	81	Fall	2
Short of breath	80	Headache	2
Cardiac arrest	54	Diarrhea	2
Referral for ECG	26	Referral: Suspected Stroke	2
Chest complaint (not pain)	22	Pneumonia	2
Abdominal pain	16	Leg swelling	2
Arm complaint	14	Sickle cell complaint	2
Elevated blood pressure	13	Mass	2
Loss of consciousness	13	Referral: Suspected Aortic Dissection	2
Weakness	10	Car accident	2
Dizziness	8	Pain (nonspecific)	1
Rapid heart beat	8	Referral: Abnormal labs	1
Weakness on one side	6	Fracture or dislocation	1
Nausea or vomiting	6	Pelvic complaint	1
Assault	5	Leg complaint	1
‘Evaluation’	5	Penile or testicular complaint	1
Ear, nose, or throat complaint	4	Cellulitis	1
Foot or ankle complaint	4	Low blood pressure	1
Fever or chills	4	Anxiety	1
Unresponsive	4	Abdominal complaint (not pain)	1
Swelling	4	Overdose	1
Device complaint	4	Flank or side pain	1
Face complaint	3	GI bleeding	1
Shoulder complaint	3	Asthma	1
Elbow, wrist, or hand complaint	3	COPD	1
Back pain	3	Referral: Suspected pulmonary embolus	1
Altered mental status	3	Blood in urine	1
Cough	3	Wound or laceration	1
Numbness	3	Referral for neurological evaluation	1
Bleeding	2	Elevated blood sugar	1
Nose bleed	2	Atrial fibrillation	1
Neck complaint	2		

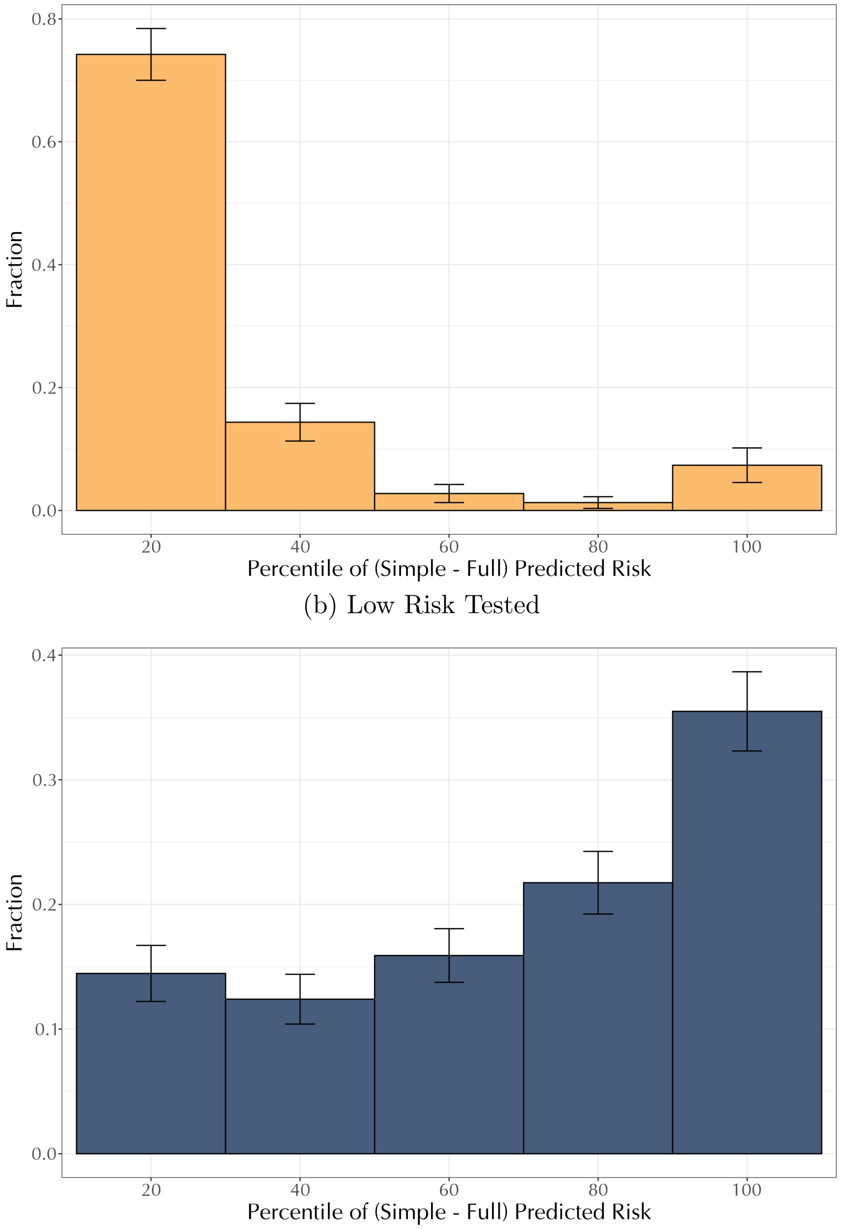
the top quintile of representativeness risk (relative to true risk) make up 34.3% come of the low-risk tested; while 99.4% of the high-risk untested come from the bottom quintile.

Table A.18: Symptom-Based Risk Scores and Yield of Testing

	Yield				
	(1)	(2)	(3)	(4)	(5)
Predicted Risk, Full	1.009***	0.946***	1.009***	0.960***	
	(0.084)	(0.118)	(0.147)	(0.124)	
Predicted Risk, Subsets					
All Symptoms		0.174	0.130	0.531	
		(0.154)	(0.174)	(0.328)	
Representative Symptoms				−0.430	
				(0.403)	
Demographics			0.180		
			(0.229)		
Prior Diagnoses			−0.044		
			(0.159)		
Prior Procedures			−0.181		
			(0.190)		
Prior Lab Results and Vital Signs			−0.101		
			(0.179)		
Physician Experience					
Experience (years)					−0.001
					(0.001)
Experience × Risk					0.015
					(0.009)
Constant	−0.001	−0.017	0.004	−0.009	
	(0.009)	(0.015)	(0.045)	(0.018)	
Observations	1,783	1,783	1,783	1,783	1,608
R^2	0.205	0.207	0.209	0.208	0.210

Note: *p<0.1; **p<0.05; ***p<0.01

Figure A.13: Distribution of Bias by Quintiles of Risk Simplicity

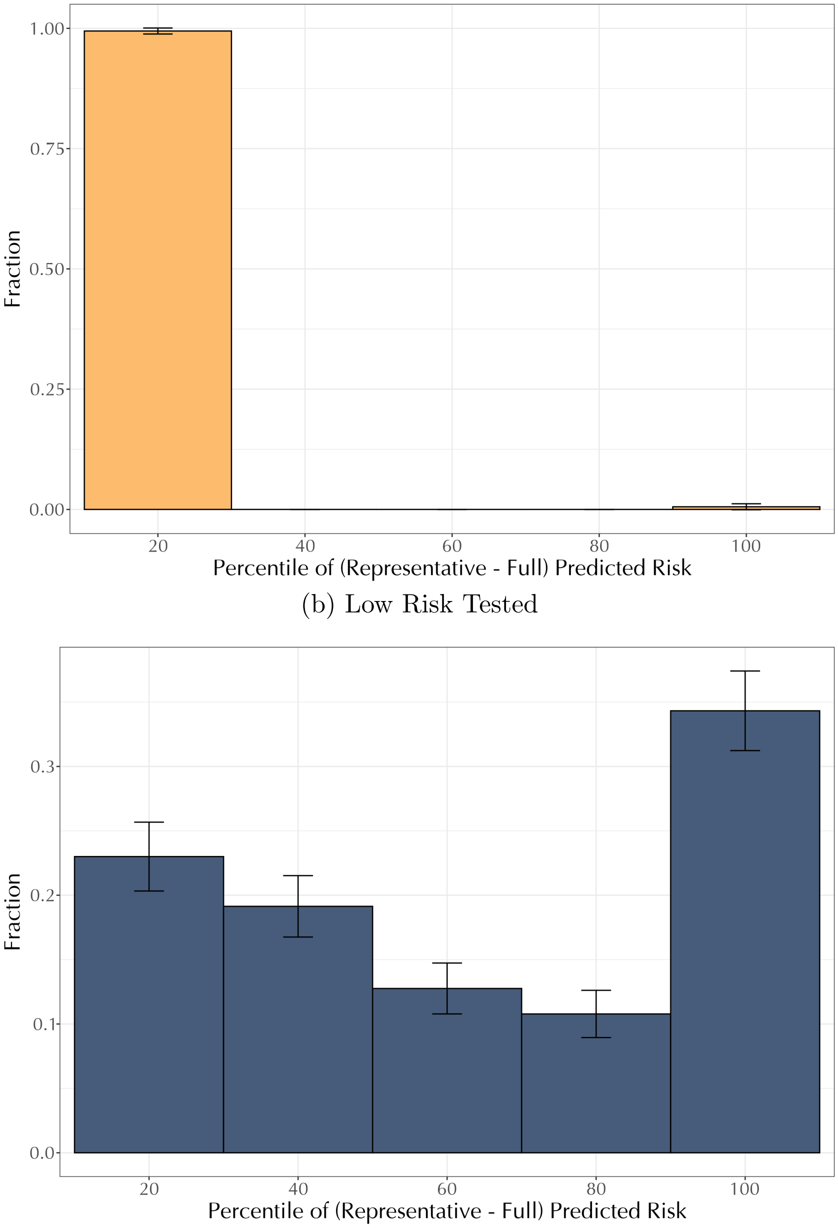


(a) High-Risk Untested

(b) Low Risk Tested

Figure A.14: Distribution of Bias by Quintiles of Risk Representativeness

(a) High-Risk Untested


(b) Low Risk Tested

10 Appendix References

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh (2016). “The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care”. In: *American Economic Review* 106.12, pp. 3730–3764.

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

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

Athey, Susan and Guido W. Imbens (2019). “Machine Learning Methods That Economists Should Know About”. In: *Annual Review of Economics* 11, pp. 685–725.

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

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

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

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

CMS, (Center for Medicare and Medicaid Services) (2016a). *Medicare Fee-for-Service Payment Acute Inpatient PPS.*

— (2016b). *Physician Fee Schedule – 2016 National Physician Fee Schedule Relative Value File.*

Eisenberg, Mark J., Kristian B. Filion, Arik Azoulay, Anya C. Brox, Seema Haider, and Louise Pilote (July 2005). “Outcomes and Cost of Coronary Artery Bypass Graft Surgery in the United States and Canada”. In: *Archives of Internal Medicine* 165.13, pp. 1506–1513.

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

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

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

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

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

Hill, J. D., J. R. Hampton, and J. R. Mitchell (Apr. 1978). “A randomised trial of home-versus-hospital management for patients with suspected myocardial infarction”. In: *Lancet* 311.8069, pp. 837–841.

Hong, Keun-Sik and Jeffrey L. Saver (Dec. 2009). “Quantifying the Value of Stroke Disability Outcomes: WHO Global Burden of Disease Project Disability Weights for Each Level of the Modified Rankin Scale”. In: *Stroke; a journal of cerebral circulation* 40.12, pp. 3828–3833.

Kleindorfer Dawn, Lindsell Christopher J., Brass Lawrence, Koroshetz Walter, and Broderick Joseph P. (Mar. 2008). “National US Estimates of Recombinant Tissue Plasminogen Activator Use”. In: *Stroke* 39.3, pp. 924–928.

Kohn, Linda T., Janet Corrigan, and Molla S. Donaldson, eds. (2000). *To err is human: building a safer health system.* Vol. 6. Washington, D.C.: National Academy Press.

Krasuski, Richard A, L Howard Hartley, Thomas H Lee, Carisi A Polanczyk, and Kirsten E Fleischmann (Jan. 1999). “Weekend and Holiday Exercise Testing in Patients with Chest Pain”. In: *Journal of General Internal Medicine* 14.1, pp. 10–14.

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

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

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

Mills, Nicholas L., Antonia M. D. Churchhouse, Kuan Ken Lee, Atul Anand, David Gamble, Anoop S. V. Shah, Elspeth Paterson, Margaret MacLeod, Catriona Graham, Simon Walker, Martin A. Denvir, Keith A. A. Fox, and David E. Newby (Mar. 2011). “Implementation of a sensitive troponin I assay and risk of recurrent myocardial infarction and death in patients with suspected acute coronary syndrome”. In: *JAMA* 305.12, pp. 1210–1216.

Miotto, Riccardo, Li Li, Brian A. Kidd, and Joel T. Dudley (2016). “Deep patient: an unsupervised representation to predict the future of patients from the electronic health records”. In: *Scientific reports* 6.26094.

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

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

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

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

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

Rajkomar, Alvin, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, and Mimi Sun (2018). "Scalable and accurate deep learning with electronic health records". In: *NPJ Digital Medicine* 1.18.

Rajpurkar, Pranav, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng (2017). "Cardiologist-level arrhythmia detection with convolutional neural networks". In: *arXiv preprint arXiv:1707.01836*.

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

Rydman, Robert J., Miriam L. Isola, Rebecca R. Roberts, Robert J. Zalenski, Michael F. McDermott, Daniel G. Murphy, Madeline M. McCarren, and Linda M. Kampe (Apr. 1998). "Emergency Department Observation Unit Versus Hospital Inpatient Care for a Chronic Asthmatic Population: A Randomized Trial of Health Status Outcome and Cost". In: *Medical Care* 36.4, pp. 599–609.

Schulman, K. A., J. A. Berlin, W. Harless, J. F. Kerner, S. Sistrunk, B. J. Gersh, R. Dubé, C. K. Taleghani, J. E. Burke, S. Williams, and others (1999). "The effect of race and sex on physicians' recommendations for cardiac catheterization." In: *The New England journal of medicine* 340.8, pp. 618–626.

Schwartz, Aaron L., Bruce E. Landon, Adam G. Elshaug, Michael E. Chernew, and J. Michael McWilliams (July 2014). "Measuring Low-Value Care in Medicare". In: *JAMA Internal Medicine* 174.7, pp. 1067–1076.

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

Sheffield, Kristin M., Patricia S. Stone, Jaime Benarroch-Gampel, James S. Goodwin, Casey A. Boyd, Dong Zhang, and Taylor S. Riall (2013). "Overuse of preoperative cardiac stress testing in medicare patients undergoing elective noncardiac surgery". In: *Annals of surgery* 257.1, pp. 73–80.

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

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

Sun, Benjamin C., Heather McCreath, Li-Jung Liang, Stephen Bohan, Christopher Baugh, Luna Ragsdale, Sean O. Henderson, Carol Clark, Aveh Bastani, Emmett Keeler, Ruopeng An, and Carol M. Mangione (Aug. 2014). "Randomized clinical trial of an emergency department observation syncope protocol versus routine inpatient admission". In: *Annals of Emergency Medicine* 64.2, pp. 167–175.

Tan, Alai, Yong-Fang Kuo, and James S. Goodwin (Sept. 2013). "Predicting Life Expectancy for Community-dwelling Older Adults From Medicare Claims Data". In: *American Journal of Epidemiology* 178.6, pp. 974–983.

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

Taylor, T. N., P. H. Davis, J. C. Torner, J. Holmes, J. W. Meyer, and M. F. Jacobson (Sept. 1996). "Lifetime cost of stroke in the United States". In: *Stroke* 27.9, pp. 1459–1466.

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

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

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

Widimský, P., T. Budešínský, D. Voráč, L. Groch, M. Želízko, M. Aschermann, M. Branny, J. Št'ásek, and P. Formánek (Jan. 2003). "Long distance transport for primary angioplasty vs immediate thrombolysis in acute myocardial infarction: Final results of the randomized national multicentre trial—PRAGUE-2". In: *European Heart Journal* 24.1, pp. 94–104.