Are We Over-Testing? Using Machine Learning to Understand Doctors' Decisions

Ziad Obermeyer, Sendhil Mullainathan*

May 2018

Abstract

Low-value health care—care that provides little health benefit relative to its cost—is a central concern for policy-makers. Identifying exactly which care is likely to be of low-value ex ante, however, has proven challenging. Here we apply machine learning tools to study an iconic decision, widely thought to epitomize low-value care: testing for heart attack (acute coronary syndromes) in the emergency setting. By comparing doctors' decisions to individualized, prospective risk estimates, we show that mis-prediction of risk is a major driver of low-value care, contributing to both over- and under-testing for heart attack. We find a substantial number of patients with very low model-predicted risk ex ante, whom doctors nonetheless decide to test. These tests are low yield, i.e., few patients benefit from interventions to treat heart attack afterwards. Indeed, individualized predictions show that the conventional approach to studying low-value testing—focusing on average, rather than marginal, yield—substantially understates the extent of over-use in the lowest-risk tested patients. So far, this fits with a common view of doctor behavior: over-testing because of financial incentives. But we also find evidence of a different kind of mis-prediction: untested patients at high model-predicted risk. Doctors' decisions not to test these patients do not appear to reflect private information: these patients develop serious complications (or death) at remarkably high rates in the weeks after emergency visits. By isolating specific conditions under which emergency patients are as-if-randomly assigned to doctors, we are able to minimize the influence of unobservables. These results suggest that both under-testing and over-testing are prevalent, and that targeting mis-prediction is an important but understudied policy priority. Finally, we use machine learning to discover specific factors correlated with over- and under-testing—e.g., over-reliance on coarse demographic factors for risk judgments, focusing on alternative, more available patient diagnoses—that give insight into behavioral mechanisms underlying physician errors.

*Obermeyer: Harvard Medical School and Brigham and Women's Hospital. Mullainathan: Harvard and NBER. For financial support we thank grant DP5 OD012161 from the Office of the Director of the National Institutes of Health. We are deeply grateful to Adam Baybutt, Christian Covington, Shreyas Lakhtakia, Katie Lin, Ruchi Mahadeshwar, Lia Petrose, Jasmeet Samra, and Aly Valliani for research assistance. This paper is a work in progress.

Health care consumes a large and ever-rising share of GDP in all countries, but the US is an outlier: 18% of GDP with 4% annual growth (Hartman, Martin, Espinosa, et al. 2017). Despite this level and growth of spending, Americans' health outcomes are falling short relative to other developed countries. As a result, low-value health care—care that provides little health benefit in light of its costs—has become a central concern for policy-makers, and widely-cited estimates (e.g., Committee on the Learning Health Care System in America, Institute of Medicine, Robert Saunders, et al. 2012) put the fraction of low-value care at one third of the $3.3 trillion in annual health care spending.

The dominant explanation for low-value care comes from the field of economics: bad incentives. Doctors get some private benefit from doing more, in the form of extra revenue or protection from risk. As a result, while individual doctors might deliver more or less care, on average they do more than they should. This view, often referred to as 'moral hazard,' has formed the basis for some of the most significant health policy initiatives in recent memory: a central component of the Affordable Care Act was a new Medicare payment scheme to reduce providers' incentives to deliver more care.

An under-appreciated but critical input to this view is a specific model of how doctors behave, and specifically how they predict risk. It holds that, for every patient a doctor sees, she forms a prediction on the expected benefit of testing or treating, and acts if the risk exceeds some threshold. This is an example of a 'prediction policy problem' (Kleinberg, Ludwig, Mullainathan, et al. 2015) where predictions on risk of a given condition are a key driver of payoffs, in this case from medical technologies.

The key insight of moral hazard for such prediction problems is that the threshold for action is set too low. Yes, doctors test those patients who are likely to benefit—but then they move further and further down the risk distribution, testing lower and lower-risk marginal patients who are unlikely to benefit. The existence of such 'over-use' is supported by the key empirical observation in the health policy literature: that many tests and interventions performed by doctors have, on average, low yield. When the vast majority of diagnostic tests come back negative, for example, and doctors are paid more to test more, it is easy to see incentives at work.

But the same model of doctor behavior—risk prediction, with testing above some threshold—might have very different effects, depending on the risk prediction regime used by doctors. For example, one could easily imagine a world where doctors, faced with the complex task of predicting which patients will benefit from a test, make mistakes. Even when incentives are aligned, medical decisions are difficult, and subject to a range of common errors and biases (Tversky and Kahneman 1974). Doctors might over-weight salient patient factors (e.g., anger, pain; see Bordalo, Gennaioli, and Shleifer 2012) or under-weight others (e.g., race, gender; see Pilote, Dasgupta, Guru, et al. 2007), and their bandwidth might be taxed by high volumes and limited time (Balogh, Miller, Ball, et al. 2016). At the extreme, if doctors' risk predictions were more or less random, average yield would be low (i.e., base rate). So mis-prediction, like moral hazard, might also create low-value care.

Historically, glossing over the particulars of doctors' risk prediction regime was partly a technical limitation: it would have been difficult to say much in the absence of accurate individual-level estimates of risk. But this is an increasingly tractable problem today, thanks to novel tools for risk prediction, and the availability of rich, high-dimensional electronic data. Unlike the traditional methods of economics or epidemiology, predictive algorithms from the field of machine learning are designed for out-of-sample accuracy, and capable of handling the types of data found in electronic datasets.

In this paper, we apply machine learning to study an iconic testing decision, often considered to epitomize low-value care: testing for heart attack (acute coronary syndromes) in the emergency setting. In doing so we build off the framework of Kleinberg, Lakkaraju, Leskovec, et al. 2017 for applying machine learning to study human decisions such as these testing decisions. Drawing on national Medicare claims, as well as detailed electronic health record data from a large hospital, we predict the yield of testing in patients presenting to emergency departments across the country, and develop three main findings.

1. We identify substantial amounts of *over-testing* nationally, in other words, patients with very low model-predicted risk, whom doctors nonetheless decide to test. These tests are invariably low-value: few of these patients ultimately receive the coronary interventions targeted to those found to have heart attack on their stress test. In cost-effectiveness terms, the bottom 30% of tests judged by model-predicted risk have a cost per life year at or above $150,000 per life year—far more expensive than the widely-used cost-effectiveness threshold of $100,000 per life year—while the top 30% come in at substantially less costly than this threshold. Thus the current literature's focus on average cost, $120,000 in our sample, both under-states the extent of predictably low-value testing in the marginal tested patients, and ignores the existence of decidedly cost-effective tests in the highest-risk tested patients.

2. We also find evidence of what appears to be *under-testing*, in other words, patients with very high model-predicted risk, whom doctors decide not to test. These marginal untested patients have a high predicted yield of testing; and when they go untested, they experience catastrophic cardiac adverse events at high rates. This observation fits with a growing literature on potential under-use, for example Abaluck, Agha, Kabrhel, et al. 2016, whose structural model suggests counterfactual outcome distributions compatible with both over- and under-testing for pulmonary embolism. A related strand of evidence comes from Chandra and Staiger 2007, who find that what appears to be over- or under-use of intervention in facts reflects poor choices on the part of hospitals.

   But of course, any effort to study under-testing must grapple with a basic econometric problem: we do not observe test results for untested patients. Lakkaraju, Kleinberg, Leskovec, et al. 2017 point out that this 'selective labels' problem is endemic to applications of prediction techniques to social

science data—simply knowing the predicted distributions of outcomes in the tested vs. untested is not enough. Specifically, in our context, the appearance of under-testing based on model predictions is far from definitive, because of *private information* observed by doctors but not the statistical model. Doctors' testing decisions are based not just on factors accessible to the model, but also key diagnostic tests like electrocardiograms (ECGs) and imaging studies, the complex visual nature of which has made them historically difficult to capture in statistical models. By ignoring these data, we neglect a key unobserved factor tying doctors' decisions to the yield of testing. To illustrate this, we turn to a rich electronic health record dataset that includes ECGs, as well as structured data more commonly used in statistical models (e.g., claims-like diagnoses and procedures, labs, vital signs, etc). Using a deep learning framework to 'read' ECG findings directly off biological waveform data, we show that, among those untested patients in the highest risk bin, including information from the ECG in our model revises predicted yield of testing downward by a full 25%, by on net moving patients from higher- to lower-risk bins. This finding, which is just one channel by which private information can distort conclusions from a predictive model, illustrates the scale of the unobservables problem. Existing evidence suggestive of under-testing is also consistent with doctor's having more information than the fairly limited data we measure.

3. To deal with this problem, we apply a technique called contraction (Lakkaraju, Kleinberg, Leskovec, et al. 2017,Kleinberg, Lakkaraju, Leskovec, et al. 2017). Specifically, we draw on our institutional knowledge of the clinical setting from which our electronic health record data are derived, to find conditions where assignment of patients to doctors is as-if random. Since doctors vary by over two-fold in their propensity to test similar patients for heart attack, we can identify the testing margin and understand doctors' current testing regimes. We find that, when doctors test more or less, they do so from *across the entire risk distribution*, not just low-risk patients. This in turn suggests the large potential gains to be had from algorithmic decision making: rather than simply getting high-testing doctors to behave like low-testing doctors, a common theme in the health policy literature, we could do better: we could cut testing rates to the level of low-testing doctors—but find 55% more patients with heart attacks. Conversely, at the same level of testing as the high-testing doctors, we could cut by 61% the number of untested patients who go on to experience adverse events. Our key finding, that marginal patients are drawn from across the risk distribution, does not seem to be particular to the hospital we study: we replicate this analysis in Medicare data by exploring testing rates across high- vs. low-testing hospitals, with similar results. In summary, doctors are not simply doing too many tests or procedures, they are doing them on the wrong people.

4. Finally, we begin to elucidate some of the behavioral mechanisms underlying these errors. We use machine learning variable selection methods

to identify those patient factors associated with doctors' testing decisions over and above algorithm-predicted risk. This approach shows that doctors seem to over-weight salient information related to patients' stated complaints; they also over-weight the component of patients' risk that comes from their demographic characteristics, and correspondingly underweight risk information contained in more complex aspects of their medical histories. Doctors also seem to over-rely on Occam's razor: they often focus on alternative, more available patient diagnoses at the expense of testing high risk patients with suspicious symptoms for heart attack.

Our results suggest that mis-prediction is a major mechanism for low-value care, which drives both over- and under-testing for heart attack. Viewing medical decisions through the lens of predicted risk produces a richer understanding of doctors' decision making, and suggests a range of solutions—related to medical education, decision aids, and even payment schemes—to prevent or disincentivize errors in decisions. It also has some important implications for current health policy initiatives. If mis-prediction is widespread, ongoing efforts to change provider incentives will have less impact on low-value care than policy-makers hope. Indeed, by forcing doctors to allocate fewer tests or procedures to the same number of patients, these efforts could add cognitive load to an already complex decision, worsening mis-prediction.

The remainder of the paper is organized as follows. Section 1 describes our data, the prediction problem we set up, and the machine learning modeling strategy. Section 2 studies yield of testing among tested patients, along with cost-effectiveness analyses. Section 3 explores the private information problem, which makes it difficult to interpret risk predictions formed in the tested group (where the outcome is observed) to untested patients (where it must be inferred), using unique features of our dataset including ECG waveform analysis. Section 4 describes our approach to studying the testing decision through the lens of natural variation in physician staffing, and quantifies the scope for better algorithmic decision-making. Finally, Section 5 begins to explore behavioral mechanisms underlying our results.

1 Prediction Problem

1.1 Medical Setting and Physician Decision Making

Heart attack is a colloquial term for an acute coronary syndrome: a realized or impending blockage in the flow of blood through the coronary arteries supplying the heart, which can cause death of a patch of heart muscle. Diagnosing and treating heart attack can be life-saving: there is a large and robust evidence base that urgent 'revascularization' procedures—opening up blocked coronary arteries, either with a flexible metal tube called a stent, or with open-heart surgery—prevents both immediate effects (e.g., sudden death from arrhythmia) and longer-term sequelae (e.g., congestive heart failure). The treatment effects here are large and incontrovertible, as demonstrated in multiple international

clinical trials in the emergency setting (as opposed to treatment effects in 'stable' coronary artery disease, i.e., patients without new symptoms that prompt emergency visits, which have been questioned in recent trials).

But diagnosing heart attack is easier said than done: even life-threatening blockages can have subtle symptoms, like a subtle squeezing sensation in the chest or even just nausea. To make matters worse, these symptoms are common in the population seeking emergency care, often caused by benign problems like acid reflux or a pinched nerve (see Swap CJ and Nagurney JT 2005). Since simple tests in the emergency setting are often unrevealing, further testing is often required: 'stress testing' the heart—subjecting it to an increased work load, by asking the patient to exert herself on a treadmill, or by administering a drug—or an invasive procedure in which a catheter is inserted directly into the coronary arteries to check for blockages. These tests have been a key part in reducing rates of missed heart attack, which in the 1980s and 1990s were substantial: anywhere from 2-11% (Pope, Aufderheide, Ruthazer, et al. 2000; Schor S, Behar S, Modan B, et al. 1976; Lee, Rouan, Weisberg, et al. 1987).

Of course, there is a tradeoff: these tests are expensive, in the thousands of dollars in both direct costs and the need for overnight observation and monitoring before testing. Some tests also carry risks related to radiation and medication exposure, from kidney failure to even cardiac arrest. Patients who have negative tests incur all these costs without any benefits, and a vast literature documents both the very low average yield of testing—often as low as 1-2%—as well as the growing costs of testing nationally (see for example Foy, Liu, Davidson, et al. 2015; Rozanski, Gransar, Hayes, et al. 2013).

We model this series of physician decisions as follows.

1. The doctor estimates the probability that a patient is having a heart attack, $\hat{h_i}$.

2. If $\hat{h_i} > \frac{B_T}{C_T}$, the threshold at which expected benefits of test T exceed costs, she proceeds with testing. (For simplicity, we here refer to stress testing and catheterization together as 'testing,' and combine direct costs of testing, indirect costs like hospitalization, and adverse events like periprocedural stroke in catheterization; a fuller discussion is in the Supplement.)

3. If the test indicates an acute or impending blockage in the coronary arteries, the patient will proceed to a revascularization procedure (again, we use this term to refer to stenting and open-heart surgery together, with more details in the Supplement.) The benefits of testing B_T of course accrue only to those who receive treatment V as a result of the test, in terms of life years B_V (unifying both longer survival and freedom from sequelae like heart failure, i.e., $B_T = (B_V|T = 1, V = 1)$.

A key assumption here concerns the relationship between predicted risk of heart attack $\widehat{h_i}$ and the benefit of testing B_V. We can write the benefit of testing for an individual patient as the difference between two counterfactual outcomes, the

world in which the patient is tested B^1_{Vi} and the world where she is not B^0_{Vi}. This can in turn be decomposed into the risk of heart attack (or, more precisely, the likelihood of the physician acting on the test result) and the individual's treatment effect τ_i:

$$B^1_{Ti} - B^0_{Ti} = \hat{h_i} \cdot \tau_i$$
$$E[B^1_{Ti} - B^0_{Ti}|X_i] = E[\hat{h_i}|X_i] \cdot E[\tau_i|X_i] + Cov(\hat{h_i}, \tau_i|X_i)$$

It is easy to see that, in the absence of covariance between $\hat{h_i}$ and $\tau_i|X_i$, the benefit of testing is monotonic in risk of heart attack. Of course, this is not necessarily the case: in particular, patients with end-stage conditions or generally poor prognoses might have a high likelihood of heart attack, but low benefit from treatment (because of side effects in the setting of general frailty, or simply their own preferences). We thus exclude these patients on the basis of data available before their emergency visits (e.g., claims for nursing home or hospice care, diagnosis codes indicating cancer, dementia, etc., following the strategy outlined in Obermeyer, Cohn, Wilson, et al. 2017).

This model implies a simple cost effectiveness calculation for testing (with a fuller accounting of individual costs, benefits, and assumptions used from the literature in the Supplement):

$$\begin{aligned} E[C] &= p(T)C_T + p(V|T=1)C_V \\ E[B] &= p(V|T)B_V \\ E[B-C] &= \underbrace{p(V|T=1)[B_V - C_V]}_{\text{revascularized: } B-C} - \underbrace{p(T)C_T}_{\text{tested: only } C} \end{aligned}$$

As a starting point for our analysis, we can compute the average cost effectiveness of tests in our sample, which is $120,079 per life year. This would be considered not cost effective at the commonly-used threshold of $100,000 (Neumann, Cohen, and Weinstein 2014), and confirms findings in the literature on the low average yield of such tests. Of course, this average number can conceal quite a bit of heterogeneity. To be precise, even this estimate of cost-effectiveness might be far too optimistic: surely the marginal patient at the doctors' testing threshold would be even lower value testing. Thus to explore this and other hypotheses about the value of testing, individual-level risk estimates are required.

1.2 Measuring the Yield of Testing

A great deal of medical research is devoted to predicting heart attack risk. So it is perhaps surprising that, while the concept of heart attack seems crisp, its actual measurement remains difficult. Most large clinical trials and prospective studies, for example, define heart attack using an adjudication process in

which a committee of clinicians judges heart attack retrospectively, by reviewing laboratory studies, imaging, electrocardiograms, and patient narratives. This makes the measurement of heart attack subjective, and requires concerted effort to label cases.

In this paper, by contrast, our measurement strategy is guided by a policy question, concerning a particular decision: whom should doctors test to maximize the benefits of testing, given the current technology for identifying those who benefit and current treatment regimes. Getting this measurement right is as important in a prediction problem as in causal inference, since algorithms can replicate and even magnify the effects of mismeasurement on the left hand side (Mullainathan and Obermeyer 2017). In this case, we do not wish to replicate an abstract clinical or statistical measure of heart attack risk, but rather measure and predict two policy-relevant prediction outcomes: which patients, when tested, proceed to benefit from testing in the form of revascularization? We design our outcome measures with this objectives in mind. First, we wish to find patients who get revascularization interventions as a result of testing. To do so, we identify, among all tested patients ($T_i = 1$), which patients ultimately benefit from the test in the form of a urgent revascularization procedure (Y_i^T).

A key challenge here is that we observe this outcome only in the tested (the 'selective labels' problem, which we discuss in more detail below). Of course, we also wish to identify high-risk patients who were not tested ($T_i = 0$), but who experienced poor outcomes that might indicate that they would have benefited from testing. This would answer a related policy question: And which patients, when untested, suffer poor outcomes that might have been prevented with earlier testing? Here we take advantage of the longitudinal nature of electronic records to identify potential sequelae of untreated heart attack. In clinical trials and cohorts, this is often defined using a basket of outcomes: subsequent diagnosis or laboratory evidence of heart attack, need for a later revascularization procedure, or cardiac arrest or sudden death. We thus replicate this outcome (Y_i^U) in untested patients.

1.3 Data

Medicare Using nationally-representative Medicare claims data, we identified 20,059,154 ED visits over a four and a half year period from January 2009 through June 2013. We excluded non-fee-for-service patients, since we do not observe their full claims history. We also excluded a number of patients whose general poor health might mandate a different approach to testing, since they might not be healthy enough to undergo—or want— treatments resulting from testing: those with a visit to a Skilled Nursing Facility in the 30 days prior to the ED visit; those with a hospice claim in the year prior; those with poor-prognosis conditions diagnosed in the year prior (e.g., metastatic cancer, dementia, in whom tests and interventions may be intentionally deferred by doctors or patients; see Obermeyer, Cohn, Wilson, et al. 2017 for additional details and rationale). We also exclude those who died in the ED (i.e., a discharge code of death), and patients diagnosed with heart attack in the ED who were ultimately

not tested, likely reflecting either a known diagnosis or a specific reason a test was not performed (e.g., patient preference, known prior test results). Summary statistics on demographics and concurrent medical illnesses, and further details on International Classification of Disease (ICD) codes and laboratory studies are in the Supplement.

For the included 12,447,376 visits, we identified which visits were followed by testing for heart attack within 10 days of visits. One major but underappreciated challenge in working with electronic health records is accurate measurement of clinical tests and outcomes: a straightforward concept like 'stress test' or 'cardiac catheterization' is represented in a range of procedure codes and test result databases. There is no straightforward way to capture these: for example, three widely-cited papers on testing for heart attack use 20 or so different codes each to measure stress testing and catheterization (e.g., Schwartz, Landon, Elshaug, et al. 2014). An additional complication is that the most commonly-used procedure coding system (Current Procedural Terminology, or Healthcare Common Procedure Coding System) is updated every year, with significant changes that can lead to major discontinuities in testing rates for the same hospital over time. We performed a comprehensive search of these coding databases, and identified 59 distinct codes for catheterization and 106 for stress test (detailed in the Supplement). As one metric of the value of these additional codes over those typically used in the literature, they accounted for 11% of test and 5% of intervention procedure codes in this dataset.

@articlesheffield$_{o}veruse_{2}013, title = Overuseofpreoperativecardiacstresstestinginmedicarepatientsunde$
$Sheffield, KristinMandStone, PatriciaSandBenarroch - Gampel, JaimeandGoodwin, JamesSandBoyd, C$
$Annalsofsurgery, volume = 257, number = 1, pages = 73, year = 2013, publisher =$
$NIHPublicAccess$

@articleshreibati$_{a}ssociation_{2}011, title = Associationofcoronary CTangiographyorstresstestingwithsubse$
$Shreibati, JacquelineBandBaker, LaurenceCandHlatky, MarkA, journal = JAMA, volume =$
$306, number = 19, pages = 2128 - 2136, year = 2011, publisher = AmericanMedicalAssociation$

Overall, we identified 605,943 tested visits with stress test (389,357: treadmill or imaging) or cardiac catheterization (261,501). This window was designed to capture both those tested immediately and those referred for urgent testing after ED visits according to current guidelines and best practices. Among the tested, we identified 111,075 who had revascularization procedures (stenting: 78,124; coronary artery bypass surgery, CABG: 34,001) in the 7 days after the initial test.

In the 11,841,433 untested patients, we identified potential complications of undiagnosed heart attack on the initial visit. In the six months after ED visits, we identified subsequent diagnoses of heart attack (using International Classification of Disease codes, and need for revascularization procedures (i.e., stenting or CABG). We also observe date of death, from Medicare records.

Electronic Health Records For a subset of our analyses below, we obtain electronic health records from a large urban hospital, and re-create analyses and predictive modeling similar to that described in Medicare above. Briefly, we ob-

tained complete data (diagnoses, procedures, laboratory studies, vital signs, ED records including complaint at visit, and electrocardiograms) on all 177,825 visits to a large, urban emergency department (ED) over a three year period from 2010-12, and applied similar exclusion criteria to those described in Medicare data above. For the included 147,953 visits, we identified which visits were followed by testing for heart attack within 10 days of visits, and identified 4,773 tested with stress test (3,105: treadmill or imaging) or cardiac catheterization (1,668). Among the tested, we identified 738 who had revascularization procedures (stenting: 651; coronary artery bypass surgery, CABG: 87) in the 7 days after the initial test.

In the 143,180 untested patients, we identified potential complications of undiagnosed heart attack on the initial visit. In the six months after ED visits, we identified subsequent diagnoses of heart attack (using International Classification of Disease codes from electronic health records, and using laboratory evidence of heart attack in measured values of the cardiac troponin biomarker) and need for revascularization procedures (i.e., stenting or CABG). We excluded 4,946 patients who had a positive troponin in the ED and an accompanying diagnosis of heart attack, on the assumption that these patients likely had a reason for which either testing or revascularization procedures were impossible (confirmed by hand-review of a sample of charts; reasons included patient or family preference, known severe coronary disease refractory to treatment, etc.).

An important caveat here is that we do not measure all tests, procedures, and outcomes after patients leave the ED: only those that happen in the network of the hospital we study. While this is perhaps less of a concern for patients whose visit led directly to testing—since positive tests are highly unlikely to lead to discharge from the hospital and potential loss to follow up—it becomes important for longer-run outcomes in untested patients. Thus the rate of poor outcomes in the untested should be considered a lower bound. As one way to compensate for this, we do obtain mortality data from the state Social Security Administration, which gives us some insight into very poor outcomes in untested patients.

We first split the sample into a training set for model development, and a hold-out set for model validation (keeping all visits by the same patient together, i.e., in the training or hold-out set, to ensure that model results were not driven by recognizing individual patients). The hold-out set included all visits by any patient who had ever had a visit between the hours of 12:00am and 10:59am; the training set was comprised of all other patients. The rationale for this method of splitting the sample are described below. We were left with 64,960 visits from 51,018 patients in the training set, and 78,047 visits from 35,759 patients in the hold-out set. All results presented below are from the independent hold-out set, to which the model was never exposed in the training process.

1.4 Machine Learning

Most risk prediction for heart attack in the medical literature focuses on a handful of clinical variables, for example elements of the medical history, certain

laboratory studies, or interpreted features of the electrocardiogram (e.g., the TIMI or HEART scores). Modern electronic health records, however, contain a vast set of other data, which we feed into a machine learning algorithm to predict risk. In this section we describe these data, as well as the machine learning methods used to ensure accurate out-of-sample prediction.

Input Features To transform raw health record data into variables usable in a prediction model, we grouped ICD-9 diagnosis and ICD-9 and HCPCS procedure codes from the three-year period ending at the time of the ED visit, in addition to demographic features, into 2,720 predictors X_i. We did not use any data from the three days prior to the ED visit to fit the predictor, to avoid any leakage of information from future claims (which are occasionally back-dated). For each potential diagnosis or procedure predictor, we created two variables, the sum over two time periods: 0-1 months (recent) and 1-36 months (baseline) prior to ED visit. We dropped variables missing in over 99.9% of the training set, leaving 2,435 predictors in the model. We assumed that a missing value for a procedure or diagnosis feature implied its absence.

Algorithm We used high-dimensional statistical methods designed to handle large sets of correlated predictors, specifically gradient boosted trees: a linear combination of decision trees (Friedman 2001). A decision tree is a function $q(X_i)$ that assigns observations to a group d, identified by indices $I_d = \{i|q(X_i) = d\}$, along with a prediction w_d for each member of that group. For each observation with value X_i, the tree assigns it to a group and makes a prediction $\hat{Y}_i = \bar{Y}_{i \in I_d}$. The goal of the tree is to repeatedly partition of the covariate space in which an outcome Y_i is homogenous, and the prediction w_d. The tree is 'grown' greedily, one partition at a time, by iterating through the universe of possible candidate binary rules (dichotomizing continuous variables at threshold values, categorical variables into indicators), to split a parent node of observations identified by I_p into left and right child nodes identified by I_l, I_r. At each step, the split is chosen to maximize information gain G, the difference in entropy $H_{i \in I_d}(Y)$ of the parent node vs. the child nodes, for a given split s.

$$\begin{aligned} G_s &= H_{i \in I_p}(Y) - [p(l)H_{i \in I_l}(Y) + p(r)H_{i \in I_r}(Y)] \\ H(Y) &= -p(Y_i = 1)\log_2 p(Y_i = 1) - p(Y_i = 0)\log_2 p(Y_i = 0) \end{aligned}$$

Rather than building one decision tree, gradient boosted trees fits a series of t trees. Each tree j is fit to minimize the residual from round $j - 1$, and the final model is simply the summed predictions of individual trees:

$$
\begin{aligned}
L(Y_i, \hat{Y}_i^j) &= L(Y_i, [\alpha \hat{Y}_i^{j-1} + f_j(X_i)]) \\
F_t(X_i) &= \sum_{j=1}^{t} \alpha f_j(X_i)
\end{aligned}
$$

In practice, a weight $\alpha < 1$ (the 'learning rate') is chosen so that the full residual from round $j-1$ is not used; there are also several other parameters to choose, for example tree depth, the fraction of rows (i.e., observations) and columns (i.e., variables) to randomly sample when fitting each tree j. Choosing these parameters can be seen a trade-off between in- and out-of-sample validity: the deeper the tree, for example, the more likely we are to find complex signal in the training data—indeed, with a deep enough tree, each observation would have its own terminal node. But of course, the more likely we are to over-fit to idiosyncrasies of these data as well, hurting performance on independent data. Values for these 'tuning' parameters that maximize out-of-sample predictive accuracy are chosen in the training process.

Training Procedure We first randomly split the sample into a training set for model development, and a hold-out set for model validation (keeping all visits by the same patient together, i.e., in the training or hold-out set, to ensure that model results were not driven by recognizing individual patients). Here we describe the training, i.e., model building, procedure.

A key machine learning insight is to use the training data not just to build the model, but also to choose the optimal tuning parameters, using cross-validation to simulate out-of-sample data within the training sample. The training data are divided randomly into k 'folds' (at the patient, not visit, level); $k-1$ folds are used to fit the model and out-of-sample performance is evaluated on the kth fold. The tuning process is then repeated k times for each random sample, such that each fold contributes to both training ($k-1$ times) and testing (once). The entire process is repeated for each parameter set in the grid of parameter combinations. A sample diagram in the Supplementillustrates 4-fold cross-validation to tune one parameter in our model, the number of trees t: While in-sample (i.e., in the 3 folds used to fit the model) performance continues to improve in the number of trees, simulated out-of-sample performance (in the 4th fold) peaks then begins to fall as the model over-fits to particularities in the data used to train the model. The final model used 3000 trees of depth 16, with a learning rate of 0.005 and random row and column samples of 50%.

Final Ensemble Model The cornerstone of our model is a gradient boosted tree, trained on a joint outcome, $Y_i = (Y_i^T | T_i = 1) + (Y_i^U | T_i = 0)$. We also fit simple linear predictors of each individual outcome Y_i^T, Y_i^U using each individual X_i. In a 3% held out part of the training sample, we performed a final step to find the optimal weighting of each of these three predictions (and two interactions between the tree-based predictor and the linear predictors) using OLS.

This weighted combination was the 'ensemble' prediction that was ultimately used in the hold-out set.

Evaluation Procedure Having produced a prediction function in the training sample, we analyze its performance in a holdout sample of randomly sampled 2,481,557 visits by 770,849 patients. All results presented below are from the independent hold-out set, to which the model was never exposed in the training process. We begin by considering performance among tested patients, in whom evaluation is straightforward because we observe the outcome of testing Y_i^T.

1.5 Describing the Model

Given the complexity of the model, a succinct summary of 'which variables are doing the work' is challenging. We can perhaps see this most clearly in the model based on rich electronic health record data, where we perform a simple linear decomposition of the model-predicted risk on each of the individual variables used to produce the estimates. A few key findings emerge. First, several of the variables that account for the most variance in $\hat{y}$ are variables that clinicians might have chosen ex ante: age, a presenting complaint of chest pain or number of prior cardiac catheterizations. Second, there are also several important variables that clinicians are unlikely to have chosen: the number of mammograms a patient had in the previous 3 years, or the minimum creatinine (an indicator of both kidney function and frailty). Finally, the included variables collectively (linearly) explain only 32% of model predictions, and no one variable—whether clinically salient or not—accounts for more than 4% of variance in $\hat{y}$, illustrating the rich interactivity and non-linearity of machine learning models.

A natural question to ask is: what do we get from this added complexity in terms of accuracy? A standard measure of performance is AUC, the area under the receiver operating characteristic curve (formally, $p(\hat{Y}_i > \hat{Y}_j | Y_i = 1, Y_j = 0)$; this is preferable to accuracy since our outcome is rare, and a model could achieve high accuracy simply by predicting $\hat{Y}_i = 0$.) AUC for this model is 0.6671 (in the holdout set) for predicting whether a given tested patient will proceed to have a revascularization intervention in Medicare data, 0.7310 in electronic health records. Logistic regressions with the usual set of medically sensible variables, for instance the Framingham heart risk score, achieves AUC of 0.5903 in the Medicare data. Such small differences in AUC can translate into economically meaningful differences in prediction: for example, if we take the riskiest 1% of patients in both models, we find only a 26.1% overlap—i.e., the models largely disagree on who the riskiest patients are (Table 2). So which model is right? Looking at patients in whom the models disagree, those in the top 1% of the machine learning model but not the logit model have a realized risk of 51.1%; those in the top 1% of the logit model but not the machine learning model have a realized risk of only 27.7%. So on a range of metrics, we can see that machine learning offers substantial predictive advantage over simpler models.

2 Analysis of Physician Testing Decisions

2.1 Yield of Testing among Tested Patients

Of course, our primary concern here is not abstract accuracy of prediction, but rather the way such predictions relate to doctors' decisions. The model allows us to generate, for each Medicare patient tested for heart attack in the hold-out set, the yield of testing as a function of $\hat{Y}_i$, algorithm-predicted individual level risk. This allows us to answer a first important question about doctors' testing decisions: do doctors' test decisions correlate with predicted risk? A trivial way to approach this question is a simple logit of testing T_i on model-predicted risk $\hat{Y}_i$, yielding a coefficient of 4.83 with a standard error of (.001); in concrete terms, doctors are over 5 times more likely to test patients in the highest ventile of model predicted risk than the lowest (we discuss the relationship between model-predicted risk and testing rate in more detail below).

We can also begin to explore the testing margin used by doctors. Figure 1 shows the relationship between testing yield and predicted risk. A first observation is that the model discriminates well between high- and low-risk patients tested by doctors: yield is monotonic in predicted risk, and the model is able to identify large groups of patients with very different risk relative to the average rate of revascularization among the tested, 17.4%. Indeed, the lowest decile of tested patients in terms of model-predicted risk had only a 6.3% revascularization rate, under half the average rate.

Figure 2 shows the implications of these rates of intervention for the cost-effectiveness of tests is striking. Applying the usual $100,000 per life-year threshold, we would be left with only the top 19.8% of tests. Moreover, the occasionally-used $150,000 threshold would eliminate the the bottom 35.1% of tests (Neumann, Cohen, and Weinstein 2014). A key point is that, by providing us with a prospective risk prediction tailored to individual patients, machine learning allows us to consider the marginal, rather than the average, yield of testing, as a function of the ex ante predicted risk. This gives us a very different picture of the problem in two ways: first, it paints a stark picture of over-use of stress testing in this population: half of tests done by doctors were predictably low yield, using data available at the time of the doctors' decision. Indeed, it suggests that the conventional approach of measuring moral hazard, by measuring average yield, actually under-estimates the extent of low-value care: the patient tested on the margin has extremely low cost-effectiveness relative to the average patient. Second, it also provides a clear pathway to policy solutions: rather than exhorting doctors to test less in general, tailored risk predictions could eventually play a role in decision-making at the individual patient level.

Results from the same prediction exercise in electronic health record data (presented in the Supplement) are broadly similar to the Medicare data, or even more extreme. To summarize, average rate of revascularization among the tested was 17%, but the lowest decile of tested patients in terms of model-predicted risk had only a 2% revascularization rate. This translated into strikingly low cost effectiveness in the bottom 40% of tests (nearly $600,000 per life year).

2.2 Predictions in Untested Patients

We have seen that the model is able to identify very low-risk patients who are nonetheless tested by doctors. An interesting corollary is that many tested patients have extremely high predicted risk: in the top decile, 34.9% of patients proceed to revascularization interventions, twice the base rate. A natural question here concerns untested patients at this same level of risk: how often are they tested by doctors? Figure 3 shows a monotone increasing relationship, illustrating visually the results of the regression above that patients are tested according to risk. But a surprising finding here is that only 9.3% of patients in the highest risk decile are tested. What can we say about these patients?

This question raises two related statistical problems. First and most obviously, we have a 'selective labels' problem: while we can easily generate model predictions for untested patients, it is more difficult to evaluate the quality of the predictions. For tested patients, of course, we observe an immediate outcome for all patients: the yield of testing. For untested patients, we must find other outcomes against which to judge predictions: adverse events and need for later procedures (which, in the case of acute symptoms of heart attack, would ideally have been provided immediately to realize the maximum benefits of treatment).

A second, deeper, problem affecting any inference about untested patients is that physicians may have private information. In other words, untested patients may differ from tested patients in ways unobserved by the algorithm, and these factors may make doctors' decision not to test them quite reasonable. Certainly, the algorithm uses a rich set of data: all prior diagnosis and procedure codes, complex quantitative patterns underlying laboratory studies and vital signs, and results of prior testing. But the physician has far more information available to her: the patient's appearance, their narrative history, and immediate test results like the electrocardiogram waveform that is perhaps the most fundamental clinical tool for diagnosing heart attack.

A rough calculation illustrates the scope of the private information problem. Consider that, among 76,561 tested patients, 15,047 (i.e. 19.65%) ultimately received prompt revascularization interventions. If the 607,293 untested patients had the same rate of revascularization as tested patients with the same model-predicted risk, there would be 112,944 (i.e. 18.60%) revascularization interventions in untested patients. This would in turn imply that doctors were currently diagnosing and treating only 11.8% of all acute heart attacks. This seems implausibly low, and suggests that more investigation is needed into the mechanisms by which physicians make use of private information in their testing decision.

3 The Problem of Private Information

3.1 A Prototypical Unobservable: The electrocardiogram

A straightforward way to see the problem is to consider the data available to the algorithm, and contrast with the data available to the clinician. The algorithm

sees multiple years of medical history, as recorded in diagnoses and procedures, as well as demographics and contextual data from geography. Some of these data may be unobserved by the clinician, if the medical records are not present in the hospital system where the ED visit occurs; and of course the algorithm can make use of these data in a way that clinicians cannot (as our results from the tested patients seem to indicate). But consider how much more data the clinician has: she can speak with the patient, ask questions, and perform a physical exam. She can also observe the results of testing in the ED—including perhaps the most fundamental test for diagnosing heart attack, the electrocardiogram. As every medical student knows, some patterns are definitive for heart attack, and some patterns make heart attack far less likely (e.g., a completely normal study).

ECGs, despite their importance for diagnosis, are not usual features of health datasets, even electronic health records; they are typically stored on completely separate data infrastructure and not included in the usual sets of variables shared with researchers. But ignoring these data can have major consequences, as can be seen in Figure 4. We obtained all ECGs linked to ED encounters, 50,042 from the training set and 27,827 from the hold-out set. In these studies, we used simple regular expression matching to determine the presence of two key findings, as noted by the cardiologist in the free-text interpretation attached to the ECG: 1) 'ST elevation,' which is a worrisome indication of full occlusion in the coronary arteries; and 2) 'Normal ECG,' a summary judgment that indicates the cardiologist found no significant abnormalities in the study. We then show physician testing rate vs. our usual risk predictor using structured data, as well as yield of testing vs. the risk predictor, breaking out patients by these two ECG features.

Two main findings emerge.

1. Physician testing decisions depend heavily on ECG features, conditional on our usual risk prediction. For example, in the highest bin of model-predicted risk, patients with ST elevation are 2.9 times more likely to be tested than those with high-risk ECGs (41.67% vs 14.18%, $p < 0.001$). Conversely, those with a normal ECG are 26% less likely to be tested (11.69% vs 15.76%, $p < 0.001$).

2. These decisions correlate to true risk: yield of testing also depends on ECG features, conditional on risk prediction using structured data. Patients with ST elevation are 2.5 times more likely to receive interventions than those with high-risk ECGs (80.00% vs 31.51%, $p < 0.001$), while those with a normal ECG are 44% less likely (20.93% vs 37.42%, $p = 0.004$), all conditional on our usual risk prediction without ECG data.

Indeed, 52% of patients in the highest-risk quintile of predicted risk did not even have an ECG performed. While some of these decisions may represent errors of omission, in the majority of cases it is likely to indicate that patients had no symptoms concerning for heart attack when evaluated in the ED.

All this suggests that, absent ECG data, we would erroneously conclude that many untested patients should have been tested, because the usual algorithmic

predictions suggest high-risk. ECG data begins to show us otherwise. It also raises the point that there are innumerable unobservables that likely have the same effect: patient appearance and narrative history, tone of voice, etc.

3.2 Deep Learning on Electrocardiogram Waveform Data

There is one potential problem with the approach to ECG data outlined above. The cardiologist's interpretation of the waveform is often set down days after the visit, as she reads ECGs in large batches (to ensure reimbursement for the ECG, which does not happen without a formal interpretation—even if it comes far too late to be used in actual decision making). This introduces the possibility that additional information, not present in the waveform but inferred from other elements of the electronic record that accompany the ECG days later, are implicitly or explicitly incorporated into the interpretation, when the cardiologist interprets the study. This would make using the text of the interpretation for prediction of a past even fraught with peril.

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

We implement a residual neural network, a variant of the standard convolutional neural network used for deeper networks (He, Zhang, Ren, et al. 2016), modeled on the architecture described by Rajpurkar, Hannun, Haghpanahi, et al. 2017 for ECG analyses. The model is described more fully in the Supplement. Briefly, the model takes as input a raw electrocardiogram (ECG) signal and outputs a set of probabilities for several outcomes describing that signal. Specifically, our data consist of observations on (X_{ijt}, Y_k). X_{ijt} is a 10 second ECG signal for patient i is sampled at 100 Hz to generate a vector with $t = 1000$ time steps for each of $j = 3$ channels, corresponding to three simultaneous records of the electrical depolarization of the heart measured at three different points on the chest (leads II, V1, V5). Y_k is one of 16 outcomes coded by cardiologists; these are key features related to heart rhythm (e.g., 'atrial fibrillation'), conduction (e.g., 'bundle branch block'), and ischemia (e.g., 'ST elevation'), ascertained using regular expression matching. We then use these features (in the training set) to independently predict risk,forming a prediction that is linked purely to the ECG, as opposed to the structured data elements in the rest of the electronic health record. [1]

[1] A natural question to ask is, why predict cardiologist-interpreted features and then predict risk, rather than predicting risk directly. There are two rationales for this. First, the number of tested patients and especially the small number of positive instances (revascularization interventions), means that sample size is a major limitation. Most efforts to fit deep learning models in the literature use sample sizes in the tens or hundreds of thousands, given the sheer number of parameters that must be learned from the data. The approach of predicting cardiologist-read features present in large numbers of ECGs is one way to circumvent this problem. Second, these features are useful in our analysis of behavioral mechanisms of error

Figure 5 shows physician testing rate vs. our usual risk predictor using structured data, with patients now also broken out as a function of their ECG-predicted risk. This shows that physician testing decisions depend heavily on features of the ECG correlating to true risk, even conditional on our usual risk prediction. For example, in the highest bin of model-predicted risk, patient with low-risk ECGs are 74.51% less likely to be tested than those with high-risk ECGs (18.94% vs 10.85% $p < 0.001$). The Figure also shows the yield of testing, by structured risk prediction and ECG-predicted risk, indicating that physicians' test decisions based on ECG risk correlate with true risk: patients with high-risk ECGs are far more likely to receive interventions after testing, again conditional on model-predicted risk (42.86% vs 21.42% $p = 0.002$).

Thus in tested patients, the doctor's testing decision provides additional signal for true risk, over and above model predictions. We can apply the same logic to the untested patients, to directly illustrate why traditional model predictions alone are insufficient to conclude that physicians are under-testing. Again using the ECG features, we now build a new model that incorporates these features alongside the structure EHR data, and show this new 'updated' risk predictor against our old predictor in Figure 6. Conditional on non-ECG-based risk, there is wide variation in risk predictions that incorporate this (previously unobservable) variable. And adding ECG information results in a large net negative reclassification of patients, a total of 25% in terms of total predicted interventions, from high to low risk: many patients in the higher quintiles are reclassified down, and only a few in the lower quintiles are reclassified up when ECG data are incorporated into the predictor (e.g., 25% fewer predicted interventions in the highest quintile, or 15% of all interventions predicted in the sample; only 11% more predicted interventions in the lowest quintile, or 0.5% of all interventions predicted in the sample).

4 What Can We Say about Untested Patients?

We provide two solutions to the problem of inferring outcomes in untested patients. The first involves careful clinical curation of outcomes in untested patients, to establish that high model-predicted risk levels correlate with adverse events linked to untreated heart attack. The second takes advantage of a natural experiment relating to variation in emergency physician staffing and testing rate.

4.1 Clinical Outcomes in Longitudinal Data

We track, using Medicare claims, the frequency with which untested patients are diagnosed with heart attack in the 180 days after their emergency visits, and the need for revascularization procedures in this window. This is similar to the

below: they let us explore the possibility that doctors are missing subtle *known* features of the tracing, as opposed to features an algorithm might detect which are as yet unknown to doctors.

'major adverse cardiac event' outcome tracked in clinical trials of cardiovascular interventions like stenting, and are shown in Figure 7. In the highest decile of predicted risk, patients experience adverse events at a rate of 10.8%, suggesting that untested patients at high model-predicted risk develop adverse cardiac events in the weeks and months after their visit at rates far higher than the base rate. Of course, while a diagnosis of heart attack is suggestive, it may not be the best indication of adverse events. After all, physicians' diagnostic thresholds can vary arbitrarily, and there are many incentives to 'up-code' visits to support billing claims—but the extent of biological damage to the heart measured by laboratory studies.

One advantage of using electronic health records datasets is that they allow us to look at very granular clinical outcomes. We can thus replicate and deepen these results in Figure 8 shows, by decile of predicted risk in untested patients, the maximum measured troponin (a laboratory test that measures death of heart muscle cells) over the six months after ED visits. The usual caveats with electronic health records apply: not only would a physician need to decide to obtain the test, but the test would need to happen in the health system network of the hospital we study. So these numbers should be viewed as providing a useful lower bound on the frequency of these outcomes. That said, the results are striking. Among patients in the highest risk decile, a full 22% have biomarkers consistent with heart attack at six months, and over a third of these have substantial elevations (i.e., cTnT ≥ 0.1); these outcomes are vanishingly rare in the lower risk deciles.

Naturally, this is not conclusive evidence: in many cases, it could be rational to adopt a wait and see strategy for diagnosis. Doing tests in the emergency or inpatient setting around a visit is costly, and referring someone for outpatient follow up over the weeks or months after a visit can be reasonable. But these delays can be costly in other ways. Studies comparing an 'early invasive' strategy (i.e., catheterization within 48 hours of arrival) to a 'conservative' strategy (i.e., at the discretion of the clinician) for patients with smaller heart attacks has shown a reduction in long-term sequelae, in the form of fewer heart attacks and lower mortality at 5 years.

We find a striking increase in one-year mortality with predicted risk, shown in Figure 9. There are also more immediate consequences: fatal arrhythmias often seen in the 4 weeks after heart attack—and indeed, Figure 10 shows that diagnoses of these arrhythmias is far more common in those in the highest risk groups—conditional on surviving to diagnosis by a doctor in the hospital's health system, again a lower bound. This suggests cardiac arrest as one biologically plausible mechanism for increased mortality in patients with high predicted risk.

4.2 Natural Variation in Testing due to Physician Staffing

Of course, this evidence is suggestive but not conclusive. Ultimately, the only way to be entirely sure that unobservables are comparable between tested and untested patients is a randomized trial. Some aspects of physician staffing in

the emergency setting, however, can come close.

A common technique in health policy research is to compare variation in care delivered by providers—doctors, hospitals, regions, etc.—with health outcomes of interest. These comparisons can be challenging, because of the same unobservables problem: even after adjustment for a variety of factors, we can never be sure that variation in a particular aspect of health care is causally linked to the outcomes.

More recently, there has been renewed interest in provider variation, but with a twist: by studying settings in which patients might be more or less randomly assigned to doctors, who vary considerably in the type and quantity of care they deliver, we might be able to causally identify the link between care and outcomes. There are many settings where we might be optimistic about pseudo-random assignment of patients to doctors, particularly doctors who work specific shifts, like hospitalists or emergency physicians. At first blush, it seems that neither patient nor doctor have the ability to choose each other, minimizing the potential for unmeasured factors to distort the relationship between care and outcomes.

Unfortunately, this assumption often breaks down when subjected to scrutiny. Here we return to the granular electronic health record data, where we have access to time-stamped information on visit times and the physician responsible for the patient in the ED. In our data, for example, despite conditioning on date and time factors of patient arrival, we find that even a summary variable as fundamental as age is highly non-randomly distributed across the different doctors in our sample (F-statistic for equality of doctor fixed effects: 2.70, $p < 0.001$, controlling for year, month, day of week, and hour of arrival), as is model-predicted risk ($F : 1.63, p = 0.002$). If patients cannot choose their emergency physician, why would this happen? One key fact here is that doctors can and do choose their patients: because of overlapping shifts, when two physicians work simultaneously in the same 'pod' of the ED, there is a constant negotiation to determine who will choose which new patient. Doctors might well have preferences for different kinds of patients—simple or complex, chest pain or abdominal pain, young or old—and these preferences are likely manifested in non-random differences in their ultimate patient populations.

One institutional detail helps us here though: during certain time windows, primarily the overnight shift as well as the hours immediately adjacent to it, there is only one doctor working. This doctor sees every patient who comes in, with no room for choice, and no discretion to delay. As a result, when we restrict to these hours (starting at 12:00am, and ending at 10:59am to leave a one-hour margin before the first afternoon-shift doctor arrives), we find that both age ($F : 1.00, p = 0.466$) and model-predicted risk ($F : 1.20, p = 0.154$) are well-balanced across doctors—while testing rate remains quite different ($\chi^2 : 80, p = 0.001$; table and figures in the Supplement).

4.3 Low- vs. High-Testing Doctors

This natural randomization allows us to do two useful things. First, we can once again inspect doctors' testing decisions vs. predicted risk—but now we can explore how high- and low-testing physicians might behave differently. A common idea in the health policy literature, grounded in the idea of 'over-testing' is that we would like to encourage high-testing doctors to behave more like low-testing doctors. After all, they both test the patients who benefit the most, in the form of high yield of testing; but the high-testing doctors additionally test a number of low-risk patients who can be expected to have low yield.

The result in Figure 11 are quite different. We see that high-testing physicians test *all patients* more, drawing their marginal patients from across the risk distribution—not just low-risk patients. For example, when seeing patients in the lowest quartile of model-predicted risk, high-testing doctors are nearly three times as likely to test (0.38 vs 0.13%, $p = 0.07$); when seeing patients in the highest quartile of model-predicted risk, they are also far more likely to test (13.2 vs. 7.9%, $p < 0.001$). This suggests that a testing regime implemented by low-testing rate doctors would, in addition to testing low-risk patients less, also result in high-risk patients being tested less as well. Reassuringly, in light of the as-if random assignment of patients to doctors under our assumptions, the decision to test provides no additional information for predicting yield of testing in the tested, or adverse events in the untested, over and above model-predicted risk (Supplement).

Second, these results also suggest a useful policy simulation. Imagine that we wished to reduce testing rates. Specifically, we wish to move from a policy regime in which we tested patients at the rate of the highest-testing quartile, $\bar{T}_{q4}$, to the rate of next-lowest-testing quartile, $\bar{T}_{q3}$. We know from our data the empirical yield of testing when doctors implement the $\bar{T}_{q3}$ testing rate: the observed revascularization rate $\bar{y}_{q3}$. Now, we can simulate what would happen if we moved from $\bar{T}_{q4}$ to $\bar{T}_{q3}$, but the algorithm chose which patients to test instead of the physician. We have one key advantage here: we know the outcomes of all the patients in Q4—so we can just drop patients the algorithms considers low risk from this set, until we get to lower-testing rate $\bar{T}_{q3}$, and calculate the new simulated discovery rate: $\bar{Y}'_{q3}$. Comparing this to the actual discovery rate can illustrate the scope for improvement in algorithmic compared to actual physician decisions.

A key assumption of this method is that doctors do not make use of unobservables to guide their testing decisions differently across quintiles. For example, imagine if high-testing doctors were better at using some unobservable—an ECG, for intance—than other doctors. Indeed, maybe this is why they test more! Their tested patients would be higher risk than other doctors' patients, so when we started dropping their low-risk patients and comparing yield with lower-testing doctors, the algorithm would start the process with an unfair leg up: it starts with higher-risk, higher-yield patients. We are able to test this: we fit a model on high-testing doctors' patients, and apply this to predict yield

in other doctors' tested patients. If unobservables differed, we would expect this procedure to over-predict yield, but the predictions are well-calibrated in all testing quintiles (Supplement).

One other potential problem here is that, while patients are pseudo-randomly assigned to doctors, this is conditional on a number of time variables. In an ideal world, we would have enough sample size such that, in each cell of time (e.g., Mondays at 11pm in Q3 2012) we would have enough patients and doctors to reassign patients from high- to low-testing doctors within the particular cell. In practice, however, sample sizes are small here, so when reassigning patients from higher- to lower-testing quartiles, we re-weight patients by the conditioning variables, such that the number of tests in a simulated time cell is reweighted to match the number of tests in an observed time cell. This reassignment respects the pseudo-random assignment of patients conditional on time factors.

Our findings suggest substantial scope for improvement, as shown in Figure 12. Moving from high-testing doctors in Q4 $\bar{T}_{q4} = 0.043$ to low-testing doctors in Q1 $\bar{T}_{q1} = 0.025$, we have 42% drop in testing. Under doctors' current testing regime, observed yield (probability of revascularization among all comers to the ED) falls by 43%. If the patients had been chosen algorithmically, according to predicted risk, we could have realized the same 42% drop in testing, but only an 11% drop in yield—a 55% improvement over the doctors ($p = 0.027$).

Similarly, testing according to algorithm-predicted risk would also reduce rates of untested patients experiencing later heart attacks. Simulations of adverse events in the untested, the other side of the coin, are shown in Figure 13. When low-testing physicians in Q1 fail to test patients, they go on to have biomarker-confirmed heart attack at a rate of 1.8%—the same rate as patients seen by high-testing physicians in Q4 who do almost twice as many tests. With the same increases in testing, by contrast, the algorithm would have tested 61% more of those patients who ultimately went on to have a heart attack in the next six months compared to the high-testing doctors in Q4 ($p < 0.001$), resulting in a (simulated) adverse event rate of 0.7% in untested patients. Thus, while there can be debate over whether the policy goal should be reductions in testing, or improvements in yield, it seems likely that changing risk prediction and test selection in any way along these lines would dominate the current testing regime.

4.4 Net Over- and Under-Testing

A final question we can ask is whether, on the whole, doctors under today's testing regime are over- or under-testing. For this exercise, we generate estimates of the cost-effectiveness of testing at the individual level, based on a patient's model-predicted risk. We then categorize patients in the hold-out set as follows:

1. Among patients whom doctors currently choose to test,
    (a) Patients in whom testing is cost-effective at some threshold λ
    (b) Patients in whom testing exceeds cost-effectiveness threshold λ

2. Among patients whom doctors currently do not test,

   (a) Patients in whom testing would be cost-effective.

   (b) Patients in whom testing would exceed cost-effectiveness threshold λ

We impose one additional constraint: to deal with the problem of unobservables, we create a testing budget for each bin of algorithm-predicted risk, based on the testing rate of the highest-testing doctor quartile. Thus, after we have included the cost-effective tests that doctors currently perform, we only add the highest-risk (and cost-effective) untested patients until we reach the maximum testing rate of the high-testing doctors. This is meant to respect a fundamental inferential limitation: we can only credibly estimate yield of testing in the kinds of patients whom doctors currently test, thanks to the natural randomization across physicians. In the set of patients whom even high-testing doctors do not test, high predicted risk may not translate into high yield of testing due to unobservables, and no natural experiments allow us to test yield.

After implementing these rules, we show the results of the simulation at different thresholds λ in Figure 14. Overall, while at very strict cost-effectiveness thresholds (<\$100,000 per life-year) the dominant effect seems to be over testing—with nearly no current tests meeting the criterion—as the threshold is liberalized to values between \$100-200,000 per life year, we see that tests in both tested and untested patients would be quite cost-effective.

4.5 Low- vs. High-Testing Hospitals and Regions

Going back to Medicare data, we find that these patterns are not unique to the single hospital we study. At a gross level, we begin by comparing hospitals by their overall quintile of testing rate among all comers to the ED. We find that hospital testing rates follow a very similar pattern: Figure 15 shows that, just like the different doctors within a single institution, different hospitals that test more or less draw their marginal patients from across the entire risk distribution.

Likewise, we can also take a new look at two other commonly cited facts about the health care system through the lens of predicted risk. First, a large amount of research has found substantial regional variation in care, with some geographies testing more or less low-risk patients, leading to similar outcomes irrespective of testing rate. Second, for-profit actors in the health care system are widely believed to test large numbers of additional low-risk patients, leading to higher costs and again similar outcomes. Figure 16 shows testing rate vs. predicted risk, now separating hospitals into hospital referral regions (the geographic unit of analysis of the Dartmouth Atlas) that test more or less; Figure 17 shows the same, separating out hospitals by ownership according to American Hospital Association data. Both show that, whether we consider regional variation or ownership, actors that test more do so across the entire risk distribution, leading to more (or less) testing of both high- and low-risk patients alike.

Of course, this is far from definitive: patients might well differ on unobservable factors across hospitals, inducing changes in test rate that might be appropriate. To explore this further, we take advantage of the fact that hospitals often have staffing constraints that lead to variations in availability of advanced testing for heart attack: for example, hospitals with an in-house capability for stress testing oten do not perform these tests on weekends; catheterization laboratories are likewise not staffed on weekends, so if physicians wish to perform emergency catheterizations they must call in the relevant staff from home, at large inconvenience and expense. As a result, we might observe substantial variation in testing rate as a function of day of week of ED visits. Specifically, since patients are often tested with stress tests or catheterization on the day after they arrive in the ED (to allow a period of monitoring and observation before testing in line with hospitals' safety policies: there is a theoretical risk that stress testing could in fact precipitate a heart attack in unstable patients), we would expect to find large drops in catheterization for patients arriving on Friday and Saturday relative to Thursdays and Sundays, respectively. Figure 18 confirms this, and also our usual pattern of seeing changes in testing rate across the entire risk distribution. (For this analysis, we restrict to hospitals with an on-site catheterization laboratory, and to patients whose home zip code is within a 10 mile radius of the hospital, to avoid including those transferred specifically to the hospital for evaluation. We find no significant imbalance on important observables like age, or on model-predicted risk by day of week, as shown in the Supplement.)

Having identified a plausibly exogenous change in testing, we can now test whether our model-predicted risk is well calibrated, in two ways. First, *on average*, we can assess whether realized outcomes match predicted risk in both patients arriving during the week (i.e., on Sunday through Thursday), and weekend (i.e.g, Friday and Saturday). Second, we can algebraically assess the realized outcomes in *marginal* patients who are tested during the week, but not the weekend: since we know the testing rate and the yield of testing in each bin of model-predicted risk, we can simply subtract out the yield we *would have observed* during the weekend from the yield we *do observe* during the week; this difference gives us the yield in the marginal patients in each bin of model predicted risk. As shown in Figure 19 the model accurately predicts realized outcomes both on average and in marginal patients, giving us reassurance that the inferences we made above concerning hospital and regional variation are valid.

5 Behavioral Mechanisms of Error

What accounts for mis-prediction? The rich data elements in electronic health records, in addition to being useful for prediction, can also begin to yield clues as to the patient factors that cause doctors to over- or under-test.

To tease these out, we develop a framework for automated discovery of errors. When human decisions deviated from algorithm-predicted risk—i.e., test-

ing those and low risk, failing to test those at high risk—we can use a simple machine learning variable selection method to identify those factors associated with deviations. Specifically, we use the LASSO in a logistic regression framework to predict the testing decision, conditional on risk. The LASSO forces $\sum |\beta_k|$ to be less than a fixed value (chosen by a tuning process), effectively forcing some coefficients to be set to zero. Importantly, we do not apply the LASSO penalty to the algorithm-predicted risk variable on the right hand side, to ensure full (i.e., the OLS coefficient) adjustment for risk. We present some preliminary findings here, but note that this is work in progress.

Table 1 shows the 20 variables with the largest LASSO coefficients for predicting physicians over- and under-testing conditional on risk, respectively (for comparison, the unpenalized OLS coefficient on $\hat{y}$ is 0.561). It is important to note that all these variables are already incorporated into the predicted risk, and optimally weighted to predict the outcome; thus non-zero coefficients here imply that they are affecting testing more than they should, given their contribution to $\hat{y}$.

A few observations stand out: first, the majority of factors affecting testing over and above risk relate to the patient's 'chief complaint': the major reason for which the patient presented to the ED as recorded by the triage nurse. Complaints associated with over-testing appear to fall into two categories: those that indicate a dramatic event occurred—e.g., reported cardiac arrest—or those that cue the physician to think about the patient's heart—e.g., an chest pain. Interestingly, many of these also appear to reflect that the patient was brought in or referred in by another provider: most of those with a complaint of 'cardiac arrest' are patients brought in by an ambulance after a loss of consciousness outside of the hospital, with a presumption (although not necessarily proof) of cardiac arrest. Likewise, 'ECG abnormality' indicates that a provider outside of the ED referred the patient in for evaluation of an abnormal finding. Conversely, complaints associated with under-testing conditional on risk

Second, the next set of factors in order of importance are indicators for age-sex-race interactions: males over 70 years old, particularly white males are tested more; females, particularly those under 50 years old, are tested less, all conditional on risk. We can inspect the way physicians treat demographic information differently from other potentially relevant risk information by performing a simple decomposition of our individual risk prediction on indicator variables for each age-sex-race group. We can then see how physicians' testing decisions vary with the resulting projection of $\hat{y}$ onto demographic information, as compared to the full $\hat{y}$. Optimally, physicians would base their testing decisions on the full predicted risk, not just the component of predicted risk attributable to demographics. However, Figure 20 shows that physicians' testing decisions covary with both full risk—and the subset of risk information linked to demographic indicators. In other words, while physicians are not over- or under-testing certain demographic groups arbitrarily, they do *over-react* to demographic risk information, as opposed to the more complex risk information encoded in other parts of the patient's history.

We have seen that physicians place large weights on patients' chief com-

plaints relative to the many other aspects of patients' histories: their prior illnesses, procedures, etc. Since these elements make up a large proportion of model predicted risk, we can identify certain aspects of patients' past medical histories that might be correlated to doctors' mis-testing decisions. To do so, we create a list of medical problems whose symptoms are highly correlated with acute coronary syndromes: specifically, using EHR data where we record both the patients' chief complaints and their ultimate diagnoses, we identify those diagnoses (e.g., chronic obstructive pulmonary disease [COPD] or asthma) whose presenting symptoms (e.g., chest pain, shortness of breath) have the highest covariance with patients whom doctors ultimately decide to test for acute coronary syndromes (additional results in the Supplement).

Figure 21 shows (using Medicare data) the relationship of testing rate to predicted risk, separating patients out by whether they have a prior history (i.e., before their ED visit) of COPD or asthma. These conditions have substantial overlap with heart disease in terms of presenting symptoms, and we use a prior history of these to proxy for a doctor's likelihood of assigning them a diagnosis of COPD on the ED visit in question. We find that, particularly in the highest-risk patients, doctors test those with COPD or asthma substantially less—particularly if they had a recent (within the previous 30 days) encounter for COPD or asthma leading up to their ED visit. This has major consequences for the likelihood of later adverse events: Figure 22 shows that prior diagnoses of COPD or asthma are also highly correlated, particularly recent diagnoses. This pattern emerges for a wide range of conditions (e.g., pneumonia, anxiety) with similar presenting symptoms to heart disease (additional results in the Supplement).

All this suggests that, when presented with a complex patient who might have more than one problem—e.g., COPD and heart disease—physicians often focus on either one or the other. This may be particularly true when a diagnosis is more available, as proxied by recency of diagnosis. Medical training in diagnosis emphasizes the importance of Occam's razor; indeed, master clinicians are often portrayed as those able to take a complex, confusing set of symptoms and distill them down into a single unifying diagnosis (see Wardrop 2008). Our results indicate that, in patients with multiple coexiting problems, this may be an error.

6 Conclusions

There is increasing evidence that moral hazard cannot explain the widespread inefficiencies observed in the health care system (Baicker, Mullainathan, and Schwartzstein 2015). By looking at marginal, rather than average, yields from medical technology, we were able to illustrate the surprising extent of over-testing—indeed, this approach exposed far more than typical studies of this topic. The same approach pointed to a perhaps more surprising conclusion: despite all the incentives for physicians to over-test insured elderly patients in the emergency setting, many high-risk patients pass through the ED without

testing for heart attack, the leading cause of death in this population.

The ability to form accurate tailored risk predictions was a key part of building this evidence. This illustrates that machine learning has an interesting role to play both in applied decision making, and in testing theories in social science (Kleinberg, Lakkaraju, Leskovec, et al. 2017): comparing idealized predictions to the actions of individual actors is a fascinating new lens through which to view human behavior in complex environments. In particular, it can expose biases and reveal errors that were previously unsuspected.

Interventions to reduce low-value care have, to date, aimed squarely at sources of moral hazard, largely around the incentives of providers; often, these interventions simply reduce global rates of reimbursement for services. These interventions have produced, at best, a mixed record of targeting low-value care. More often, as the seminal RAND health insurance experiment (Newhouse and Group 1993) and more recent work since (Brot-Goldberg, Chandra, Handel, et al. 2015) has shown, changing incentives cuts all care—not just low-value care. The ability to predict the value of a specific medical intervention for a specific person opens up new channels for targeted interventions in clinical contexts, which could nudge providers to make better decisions. Interventions that improve the practice of medicine, rather than ones that simply change the incentives to practice it in a certain way, could be a powerful policy lever to drive efficient health care use.

References

1. Hartman, M, Martin, AB, Espinosa, N, Catlin, A, and The National Health Expenditure Accounts Team. National Health Care Spending In 2016: Spending And Enrollment Growth Slow After Initial Coverage Expansions. Health Affairs 2017;37:150–160.
2. Committee on the Learning Health Care System in America, Institute of Medicine, Robert Saunders, Leigh Stuckhardt, and J. Michael Mcginnis. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. Washington: National Academies Press, 2012.
3. Kleinberg, J, Ludwig, J, Mullainathan, S, and Obermeyer, Z. Prediction Policy Problems. American Economic Review 2015;105:491–95.
4. Tversky, A and Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. Science 1974;185:1124–1131.
5. Bordalo, P, Gennaioli, N, and Shleifer, A. Salience Theory of Choice Under Risk. The Quarterly Journal of Economics 2012:qjs018.
6. Pilote, L, Dasgupta, K, Guru, V, et al. A comprehensive view of sex-specific issues related to cardiovascular disease. Canadian Medical Association Journal 2007;176:S1–S44.
7. Balogh, EP, Miller, BT, Ball, JR, and others. Improving Diagnosis in Health Care. National Academies Press, 2016. (Visited on 2016).
8. Kleinberg, J, Lakkaraju, H, Leskovec, J, Ludwig, J, and Mullainathan, S. Human decisions and machine predictions. The Quarterly Journal of Economics 2017;133:237–293.
9. Abaluck, J, Agha, L, Kabrhel, C, Raja, A, and Venkatesh, A. The determinants of productivity in medical testing: Intensity and allocation of care. American Economic Review 2016;106:3730–64.
10. Chandra, A and Staiger, DO. Productivity spillovers in health care: evidence from the treatment of heart attacks. Journal of Political Economy 2007;115:103–140.
11. Lakkaraju, H, Kleinberg, J, Leskovec, J, Ludwig, J, and Mullainathan, S. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017:275–284.
12. Swap CJ and Nagurney JT. Value and limitations of chest pain history in the evaluation of patients with suspected acute coronary syndromes. JAMA 2005;294:2623–2629.
13. Pope, JH, Aufderheide, TP, Ruthazer, R, et al. Missed diagnoses of acute cardiac ischemia in the emergency department. New England Journal of Medicine 2000;342:1163–1170.

14. Schor S, Behar S, Modan B, Barell V, Drory J, and Kariv I. Disposition of presumed coronary patients from an emergency room: A follow-up study. JAMA 1976;236:941–943.

15. Lee, TH, Rouan, GW, Weisberg, MC, et al. Clinical characteristics and natural history of patients with acute myocardial infarction sent home from the emergency room. American Journal of Cardiology 1987;60:219–224.

16. Foy, Liu, Davidson, Sciamanna, and Leslie. Comparative effectiveness of diagnostic testing strategies in emergency department patients with chest pain: An analysis of downstream testing, interventions, and outcomes. JAMA Internal Medicine 2015;175:428–436.

17. Rozanski, A, Gransar, H, Hayes, SW, et al. Temporal Trends in the Frequency of Inducible Myocardial Ischemia During Cardiac Stress Testing1991 to 2009. Journal of the American College of Cardiology 2013;61:1054–1065.

18. Obermeyer, Z, Cohn, B, Wilson, M, Jena, AB, and Cutler, DM. Early death after discharge from emergency departments: analysis of national US insurance claims data. bmj 2017;356:j239.

19. Neumann, PJ, Cohen, JT, and Weinstein, MC. Updating Cost-Effectiveness — The Curious Resilience of the $50,000-per-QALY Threshold. New England Journal of Medicine 2014;371:796–797.

20. Mullainathan, S and Obermeyer, Z. Does Machine Learning Automate Moral Hazard and Error? American Economic Review: Papers and Proceedings 2017;107:1–5.

21. Schwartz, AL, Landon, BE, Elshaug, AG, Chernew, ME, and McWilliams, JM. Measuring low-value care in Medicare. JAMA internal medicine 2014;174:1067–1076.

22. Friedman, JH. Greedy function approximation: a gradient boosting machine. Annals of statistics 2001:1189–1232.

23. He, K, Zhang, X, Ren, S, and Sun, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016:770–778.

24. Rajpurkar, P, Hannun, AY, Haghpanahi, M, Bourn, C, and Ng, AY. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. ArXiv e-prints 2017.

25. Wardrop, D. Ockham's Razor: sharpen or re-sheathe? Journal of the Royal Society of Medicine 2008;101:50.

26. Baicker, K, Mullainathan, S, and Schwartzstein, J. Behavioral hazard in health insurance. The Quarterly Journal of Economics 2015;130:1623–1667.

27. Newhouse, JP and Group, RCIE. Free for all?: lessons from the RAND health insurance experiment. Harvard University Press, 1993. (Visited on 2016).

28. Brot-Goldberg, ZC, Chandra, A, Handel, BR, and Kolstad, JT. What does a deductible do? the impact of cost-sharing on health care prices, quantities, and spending dynamics. National Bureau of Economic Research Working Paper 2015.

Figures

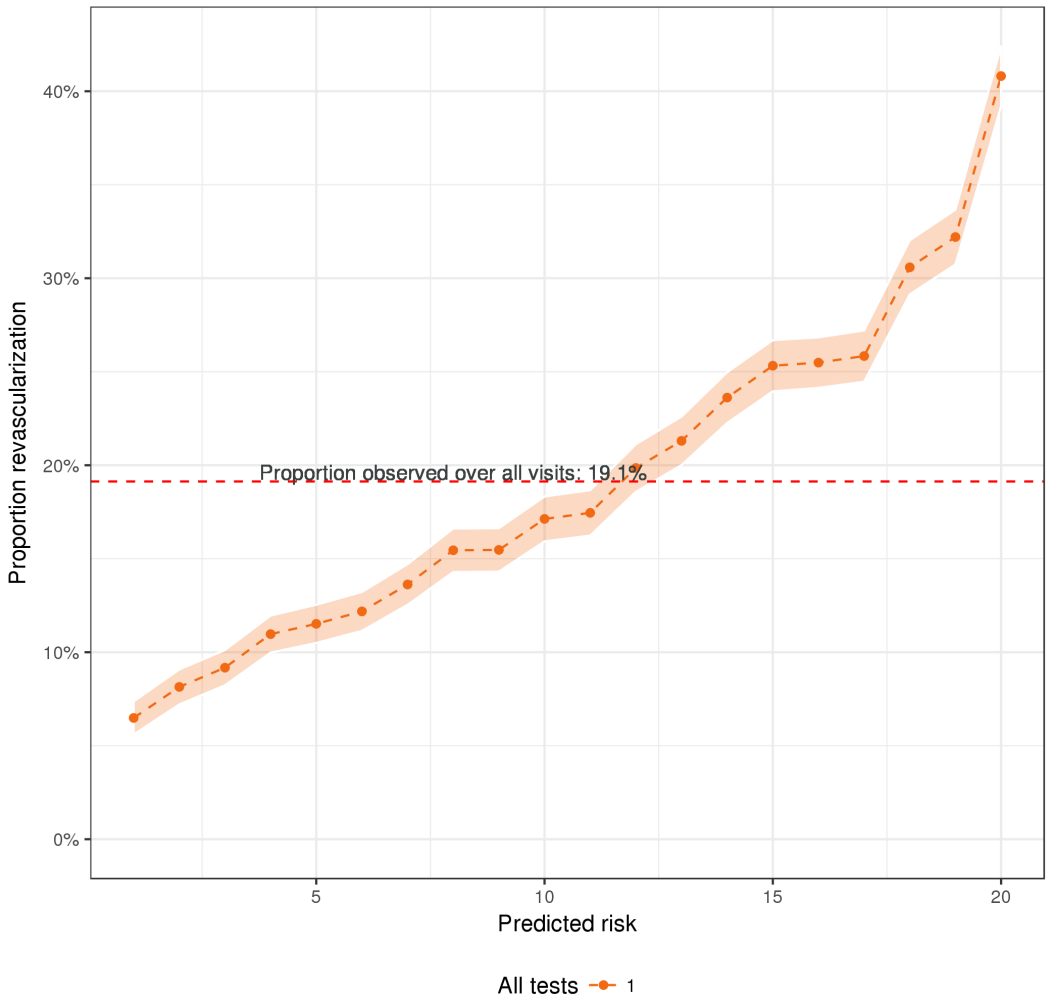


Figure 1: Yield of testing vs. decile of predicted risk among tested patients in the hold-out set. In this and all figures, the 95% CI accounts for clustering at the patient level.

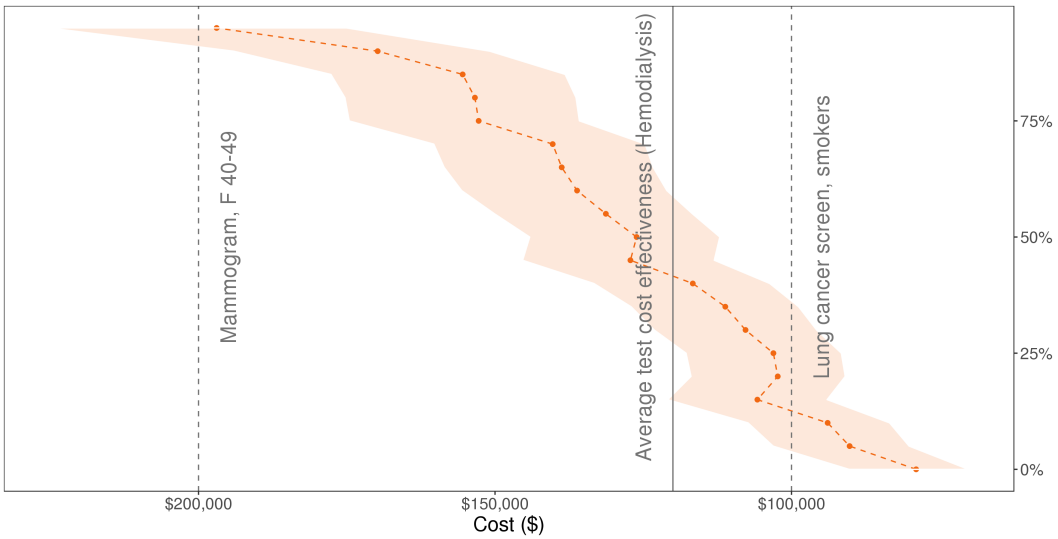


Figure 2: Cost effectiveness of tests in the hold-out set. Starting on the left of the graph, the curve shows, for each decile of model-predicted risk, the fraction of tests that would fall at or below a given cost-effectiveness threshold. For example, if we were to apply the usual $100,000 per life-year saved threshold, we would only do around 50% of all tests currently ordered by doctors.

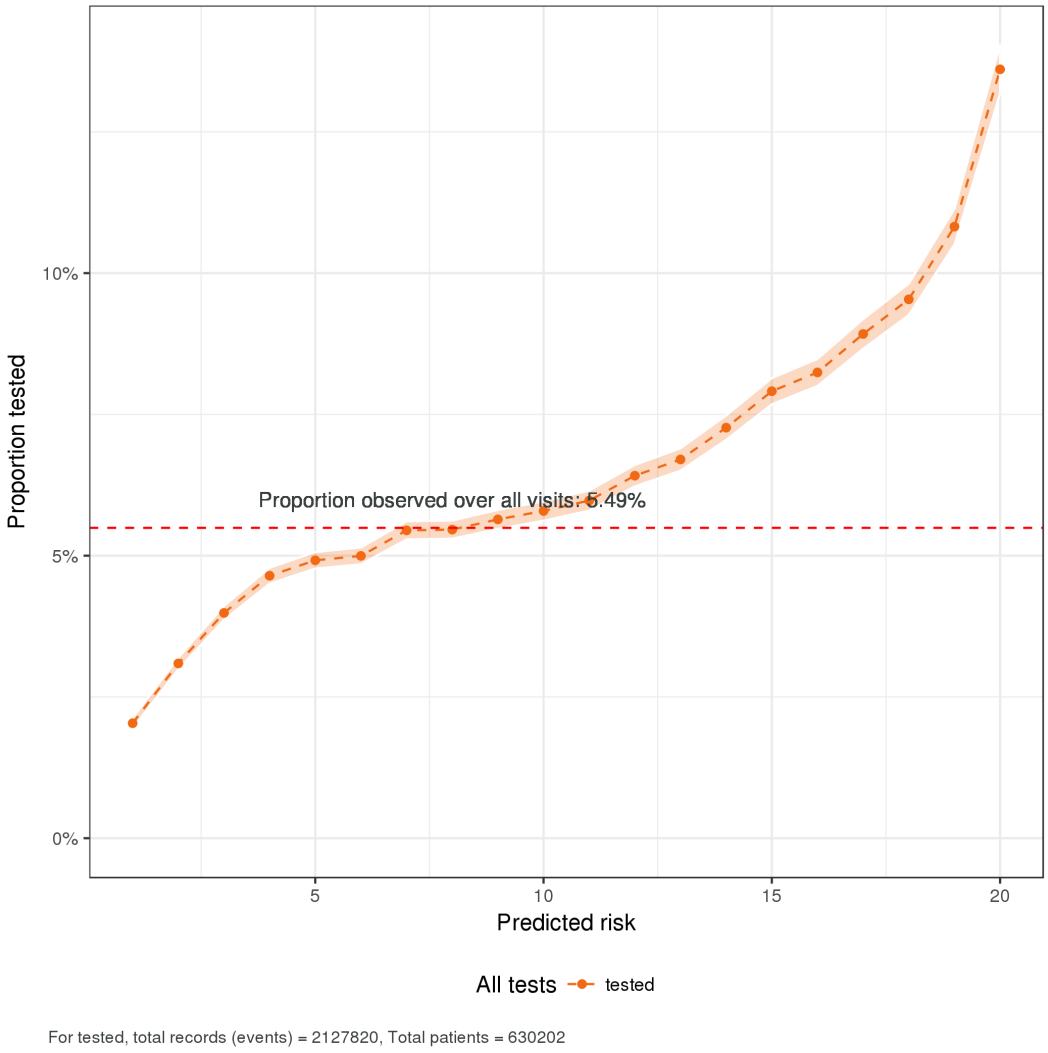


Figure 3: Rate of testing vs. decile of model-predicted risk, among all patients in the hold-out set.

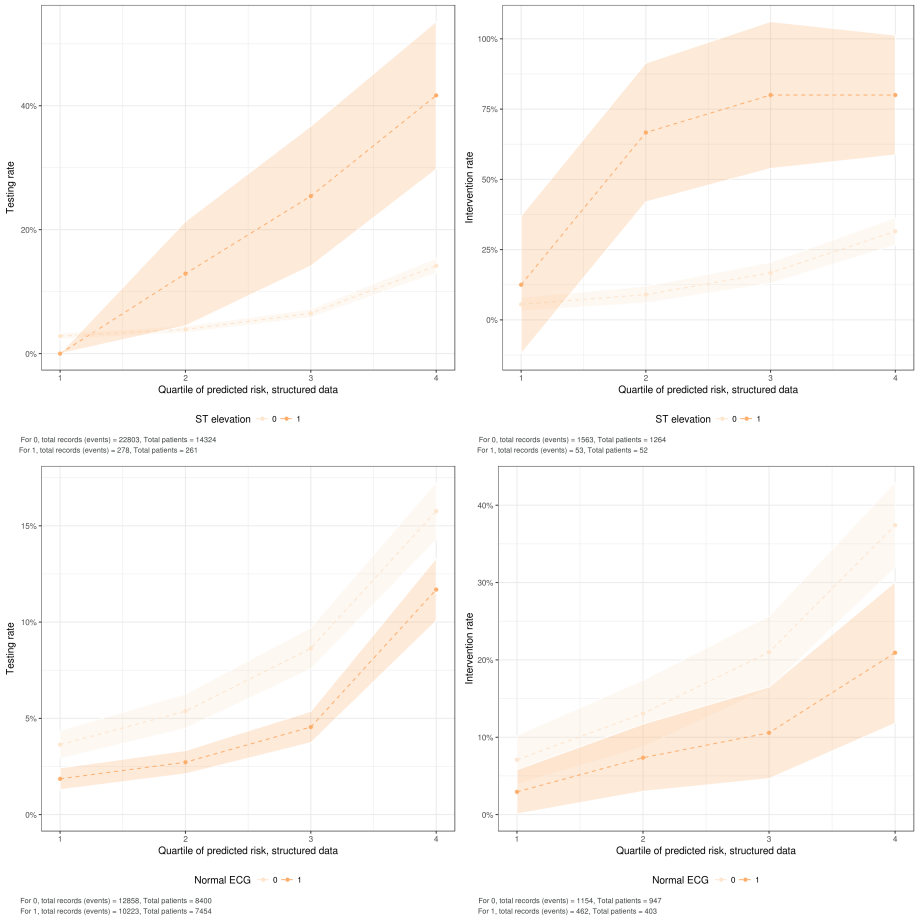


Figure 4: Rate of testing (left) and yield of testing (right), vs. quartile of model-predicted risk, by presence of absence of specific features of the ECG. The top panels show patients with and without ST-elevation, as interpreted by a cardiologist, an ECG finding indicative of heart attack; the bottom panels show patients in whom a cardiologist has judged the study 'normal', vs. all other patients with abnormalities.

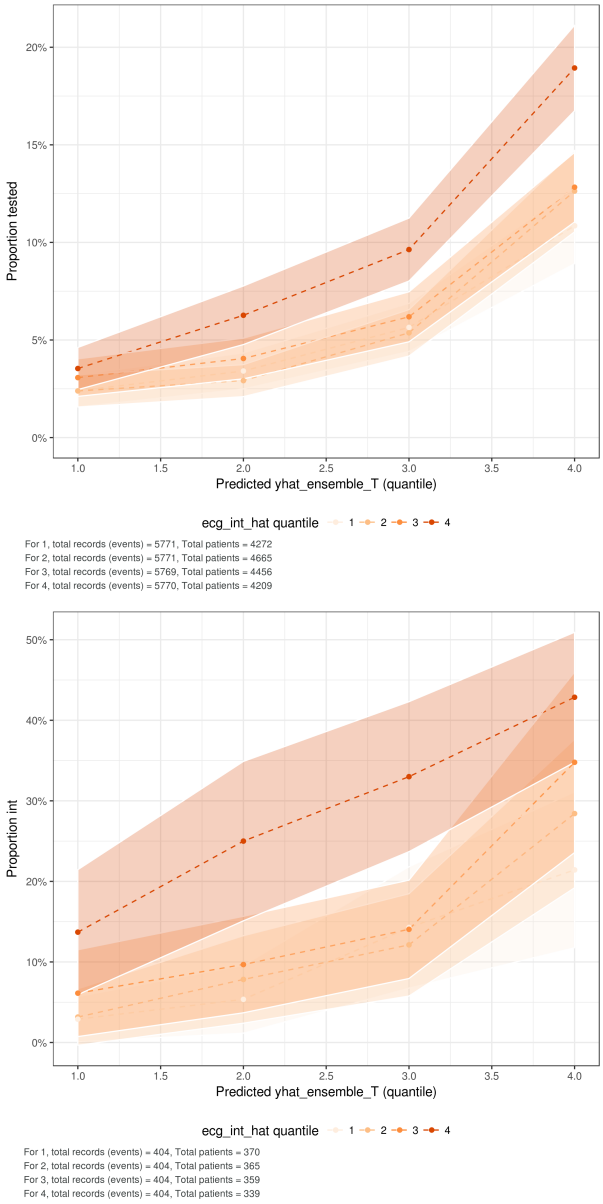


Figure 5: Rate of testing (top) and yield of testing (bottom), vs. quartile of model-predicted risk, among all patients in the hold-out set with an ECG. Patients are broken out by quartile of the ECG-based risk, using learned waveform features of the ECG (rather than cardiologist interpretations, which are often set down days after the study).

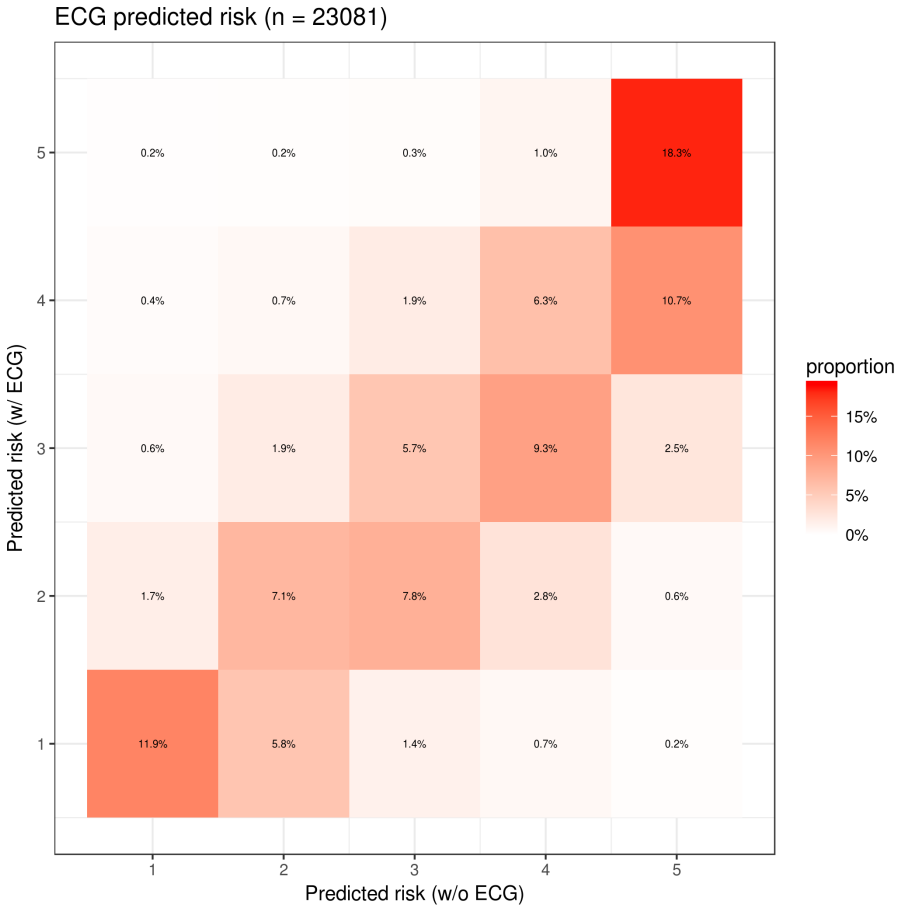


Figure 6: Comparison of model-based risk estimates, with (y-axis) and without (x-axis) incorporation of learned ECG waveform features. Bins are constructed in absolute $\hat{y}$ space on the original predictor, and preserved to bin the new predictor incorporating ECG risk information.

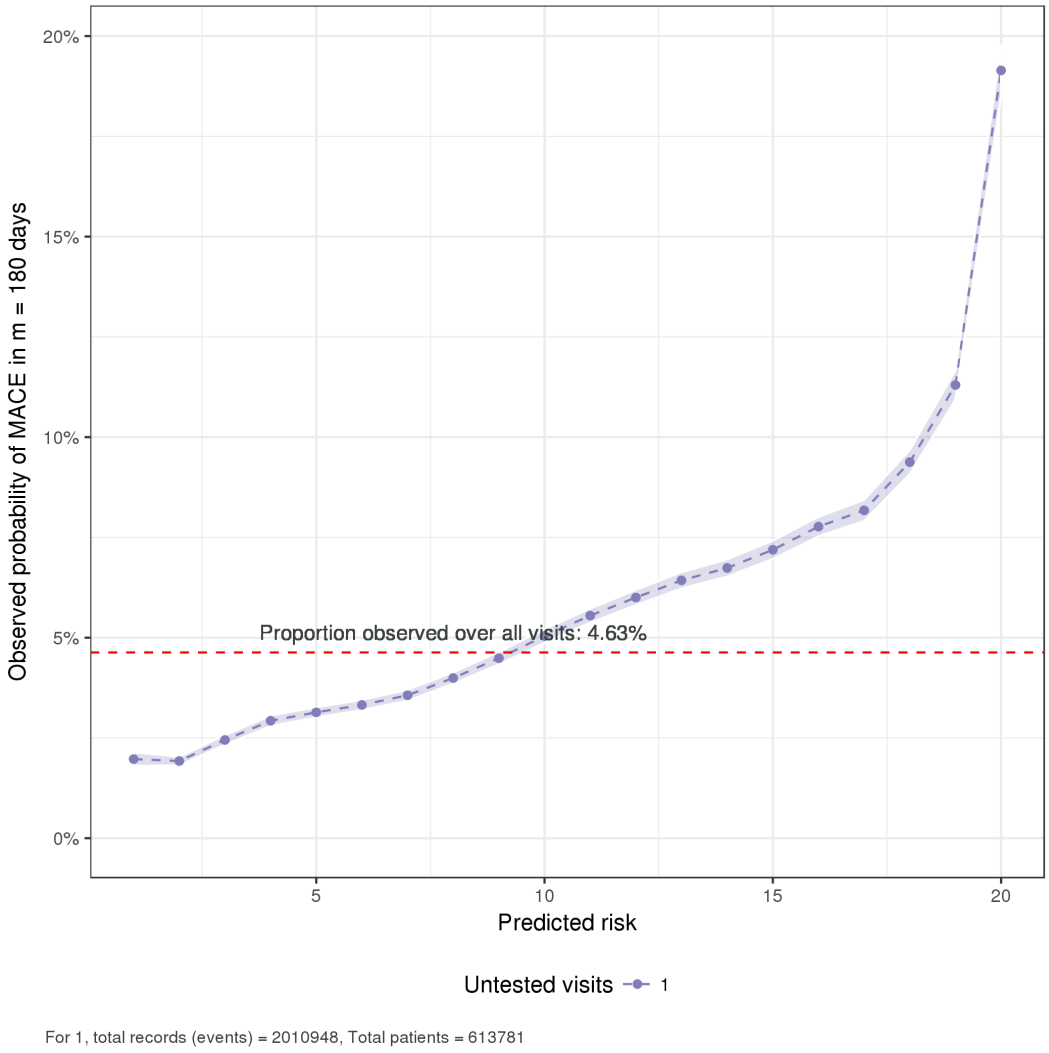


Figure 7: Rate of major adverse cardiac events (heart attack and need for revascularization), in the 180 days after emergency visits, vs. decile of model-predicted risk, among untested patients in the hold-out set.

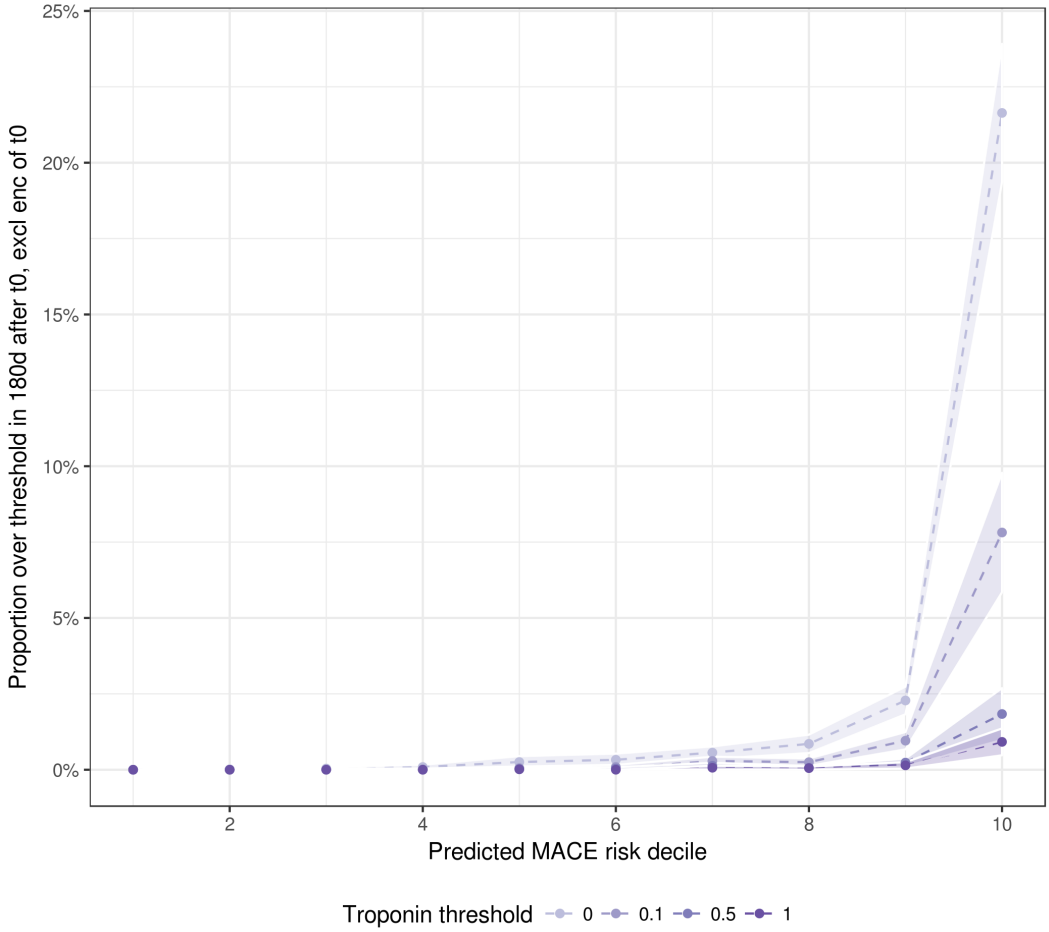


Figure 8: Rate of positive troponin, measuring damage to heart muscle, in the 180 days after emergency visits, vs. decile of model-predicted risk, among untested patients in the hold-out set.

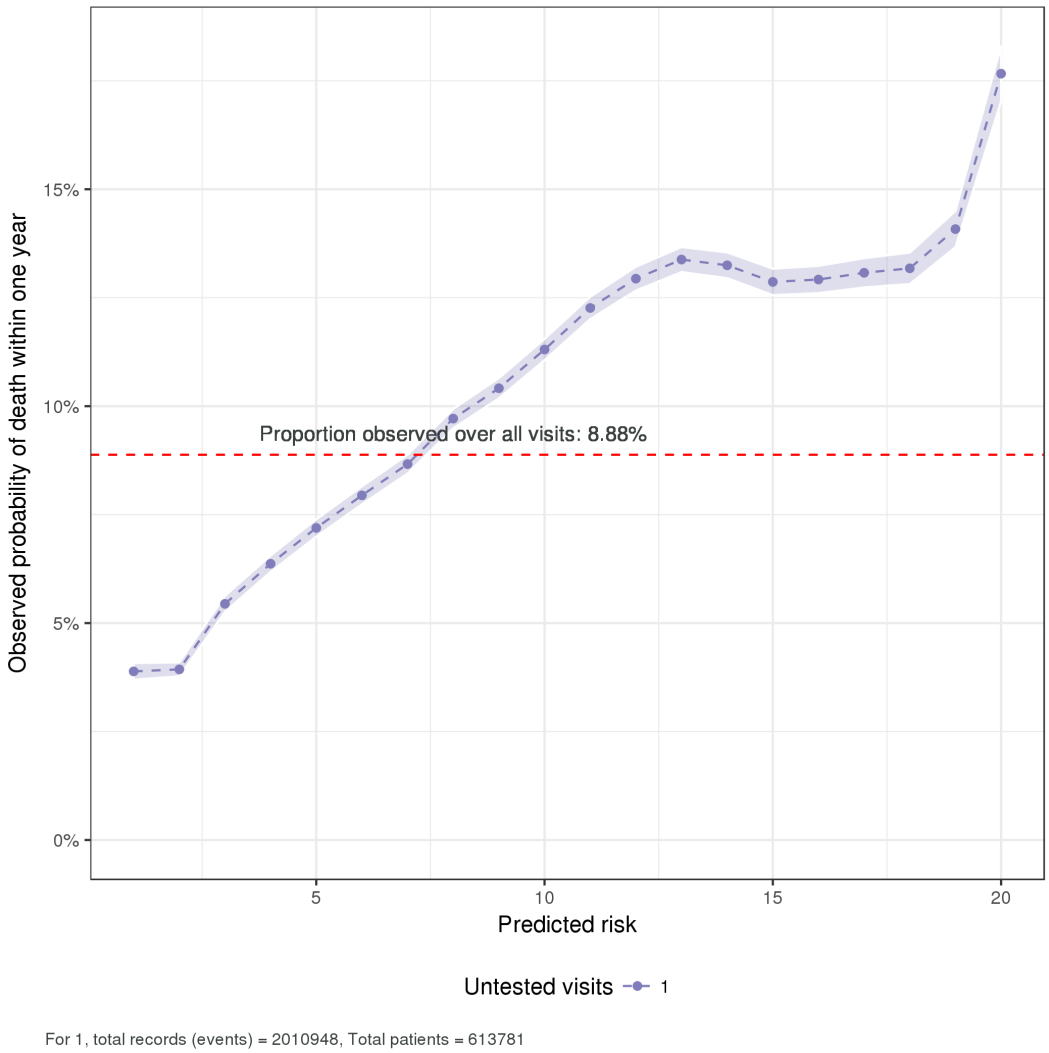


Figure 9: Rate of one-year mortality vs. predicted risk decile, among untested patients in the hold-out set.

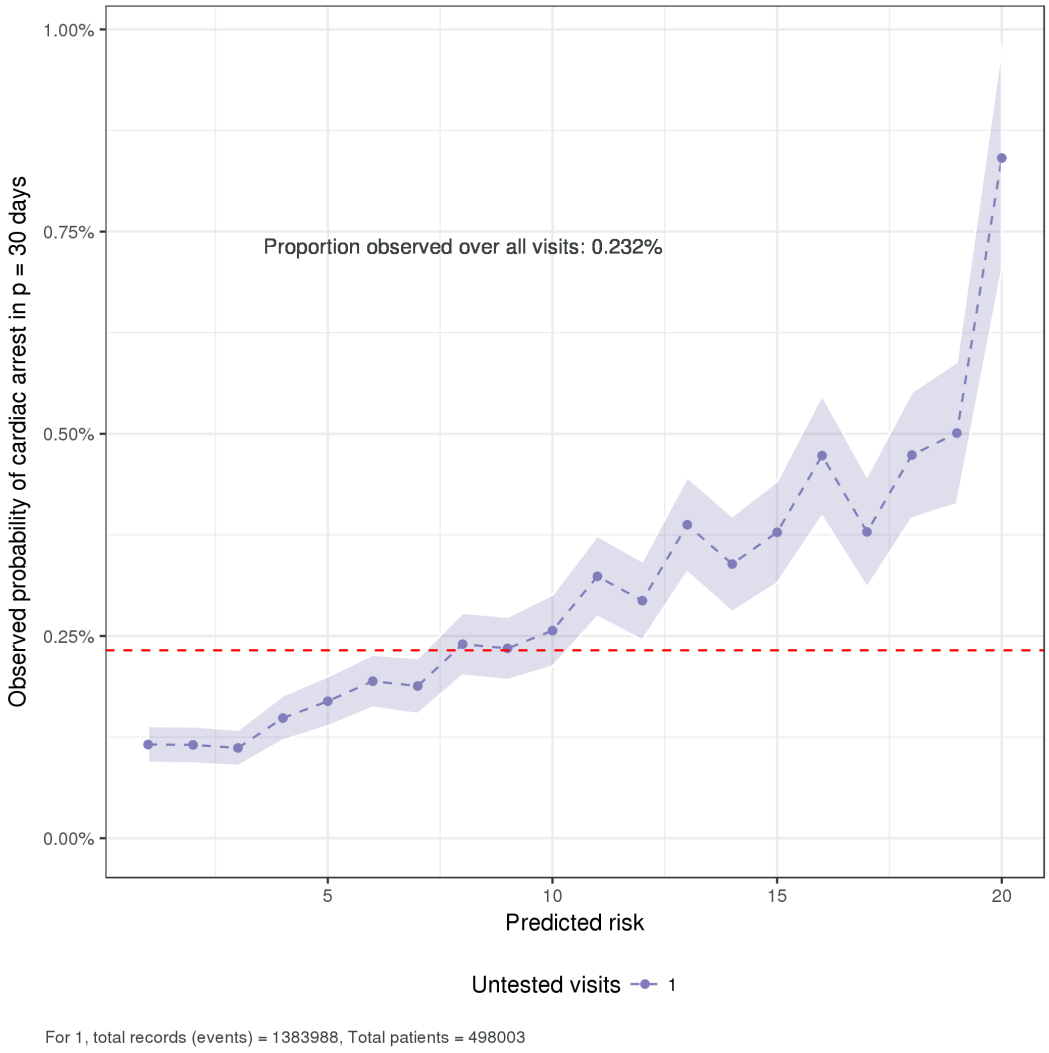


Figure 10: Rate of diagnosed cardiac arrest, in the form of malignant arrhythmias (ventricular tachycardia and ventricular fibrillation) vs. predicted risk decile, among untested patients in the hold-out set.

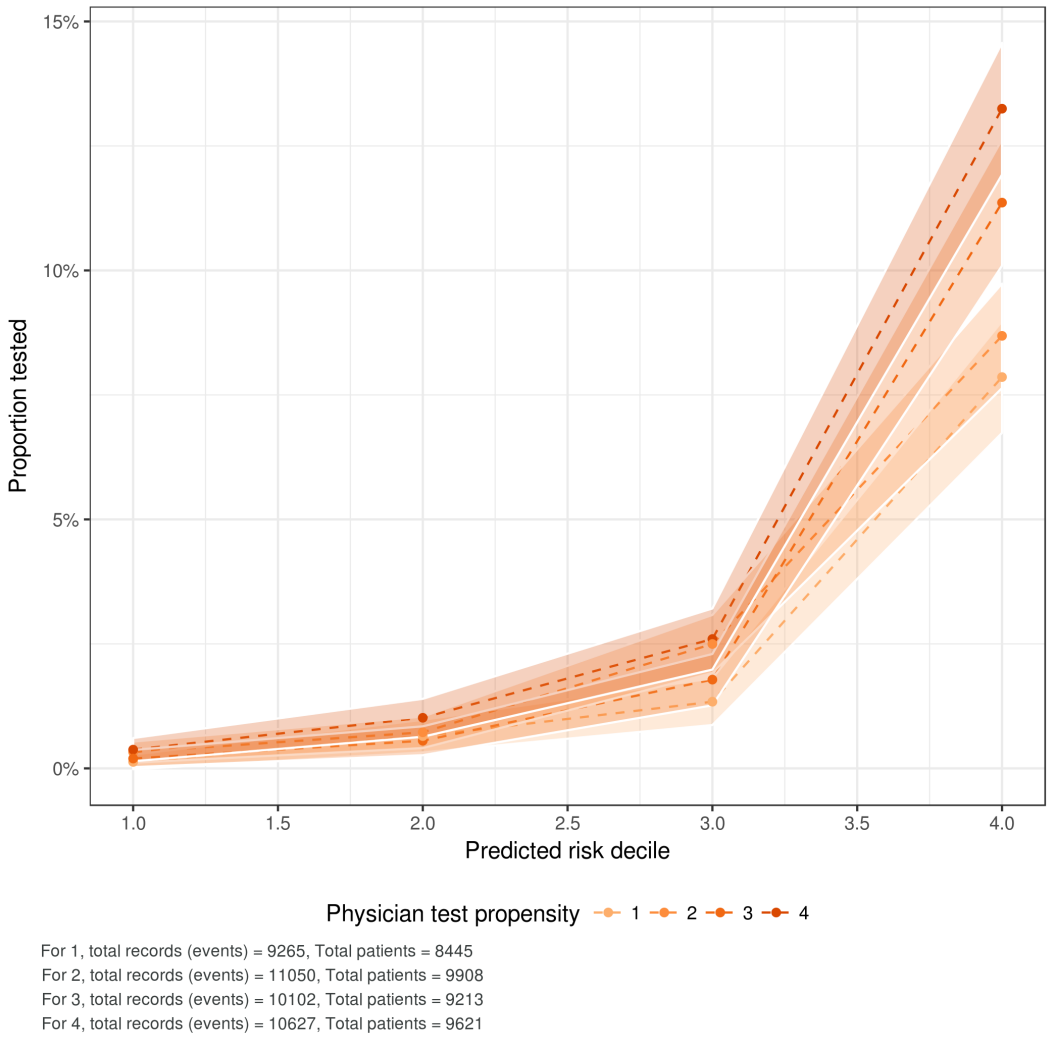


Figure 11: Testing rate as a function of model-predicted risk, by quartile of physician testing rate.

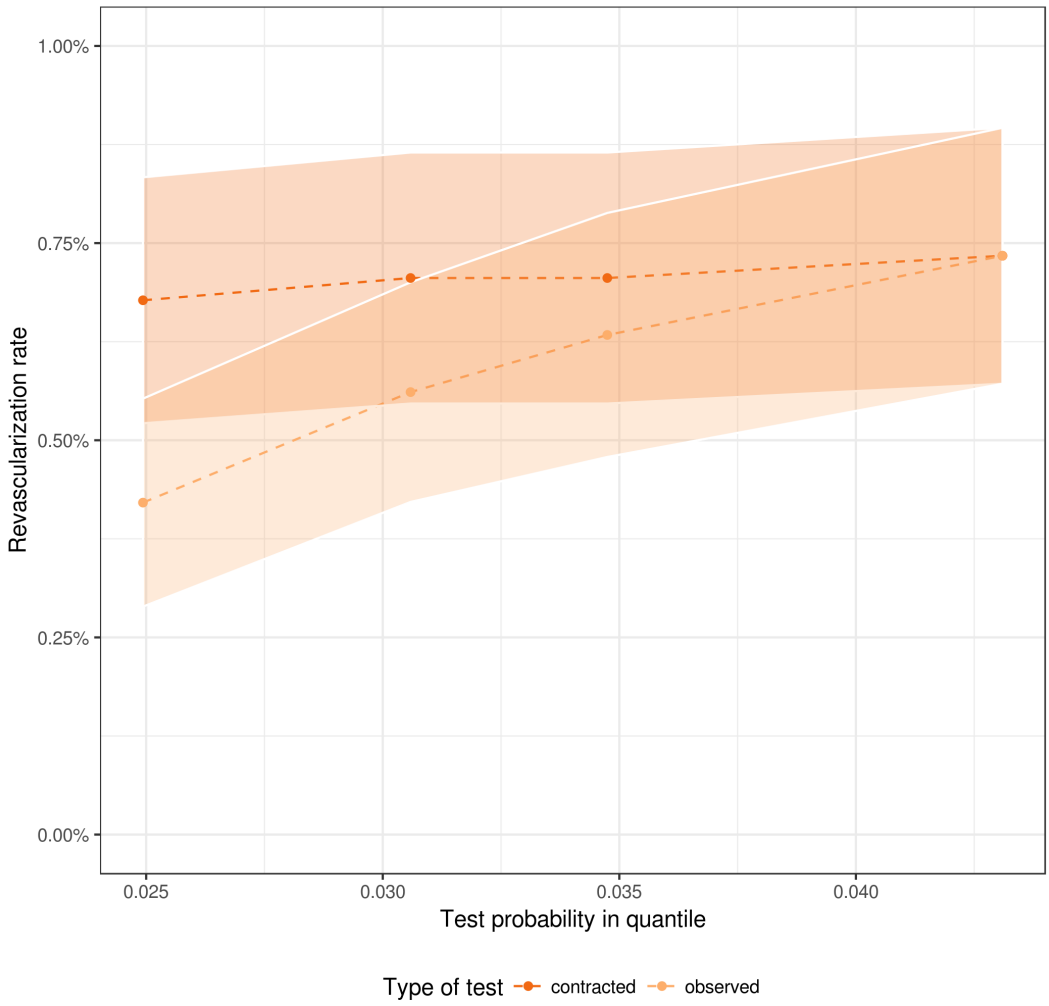
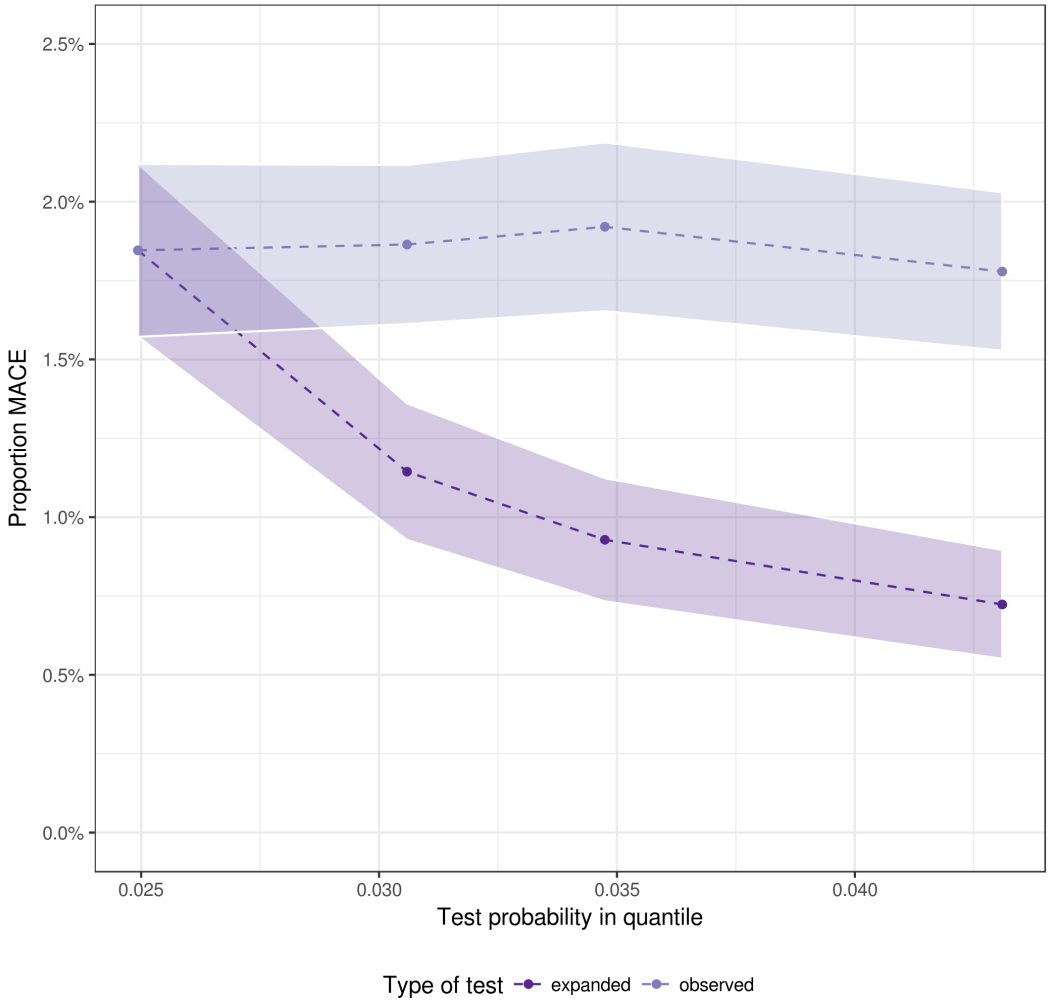


Figure 12: The solid line shows observed yield of testing in tested patients, at the testing rates on the x-axis defined by mean rate in each physician testing quartile. The dotted line shows simulated outcomes, at the testing rates of in each physician testing quartile, if patients had been selected for testing using model-predicted risk rather than physician predictions.



Welch Two Sample t-test, Q4: p-value of diff. in means = 0.000000, 95pct CI = (0.007505, 0.013602)

Figure 13: The solid line shows observed heart attack rate in untested patients, at the testing rates on the x-axis defined by mean rate in each physician testing quartile. The dotted line shows simulated outcomes, at the testing rates of in each physician testing quartile, if patients had been selected for testing using model-predicted risk rather than physician predictions.

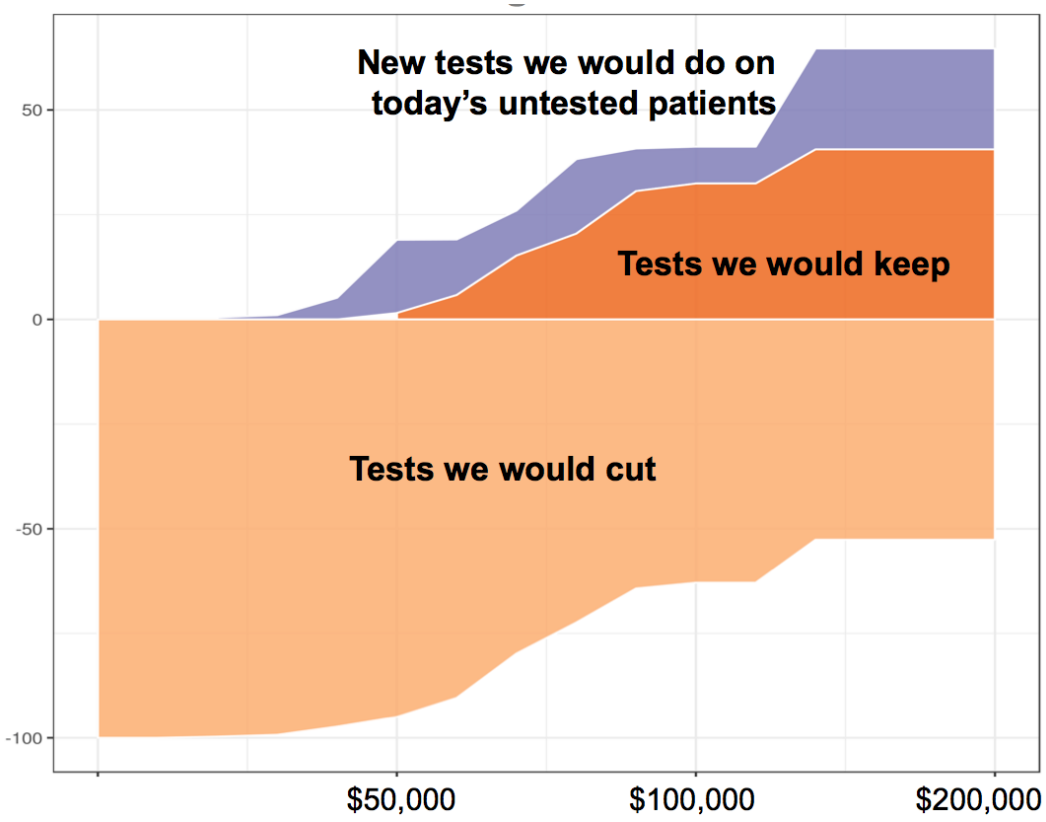


Figure 14: Simulation of cost-effectiveness of testing, by model-predicted risk, at different thresholds. The y-axis shows the percent of tests doctors currently perform that would be eliminated ('bad tests'), retained ('good tests'), and untested patients whom the algorithm would test at the testing budget ('good potential tests').

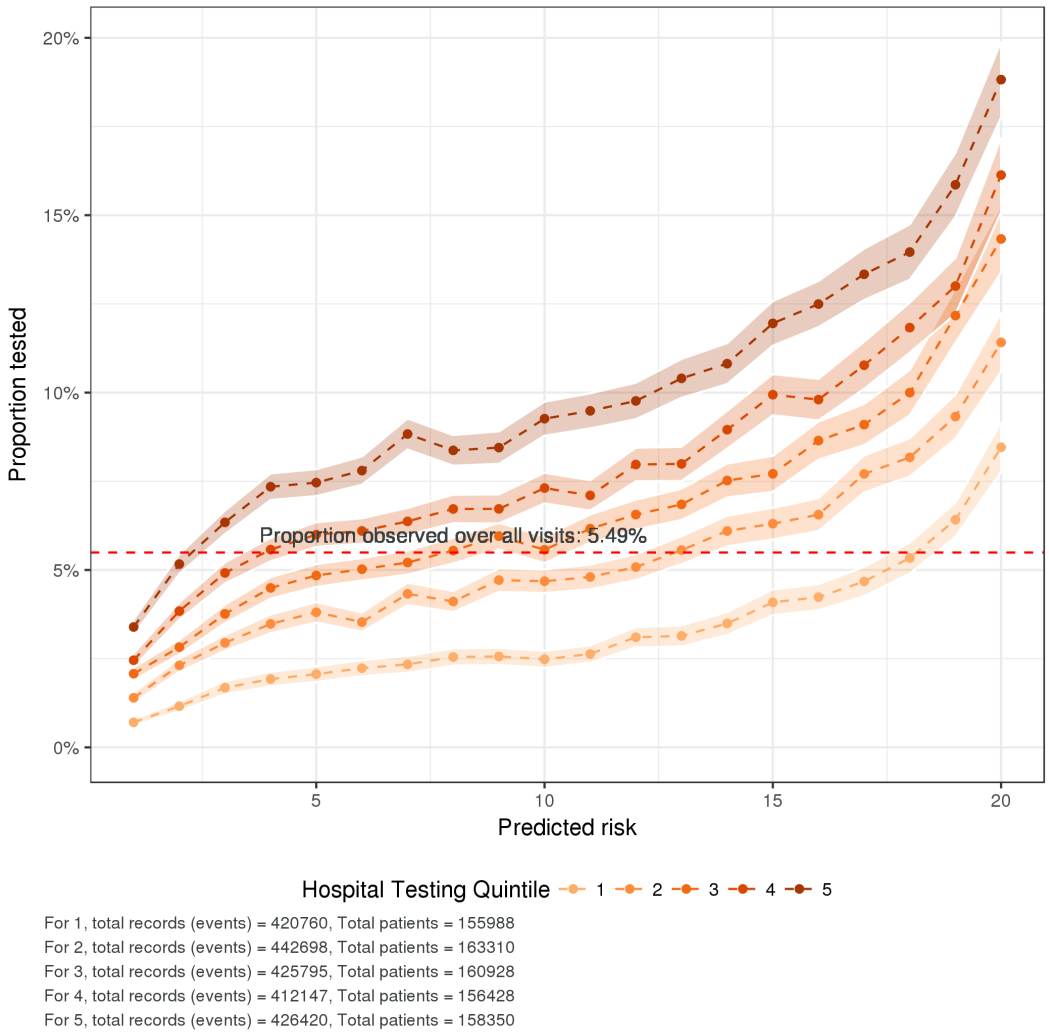


Figure 15: Testing rate as a function of model-predicted risk, by quintile of hospital testing rate.

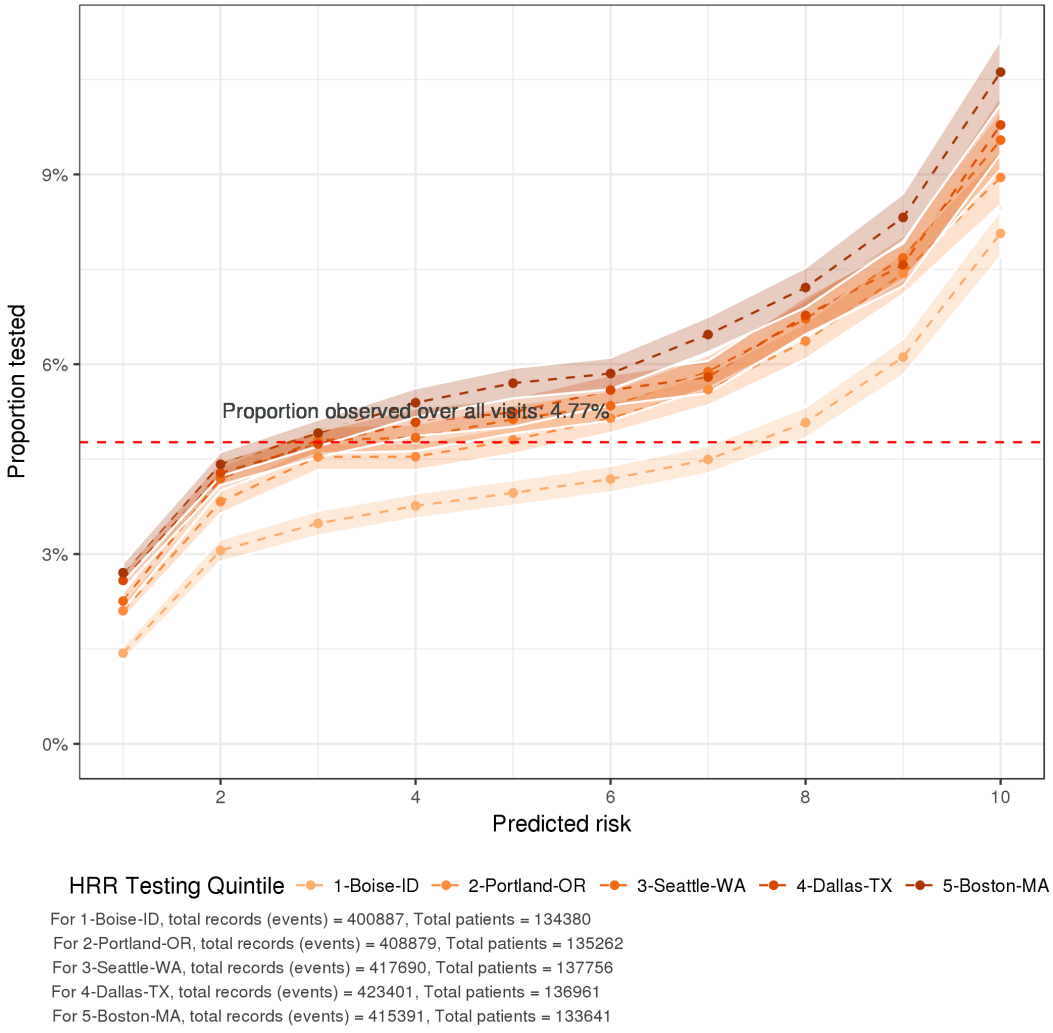


Figure 16: Testing rate as a function of model-predicted risk, by quintile of hospital referral region (HRR). Regions are labeled by the most populous single region within the quintile, e.g., Boston, MA in the top testing quintile and Boise, ID in the bottom quintile.

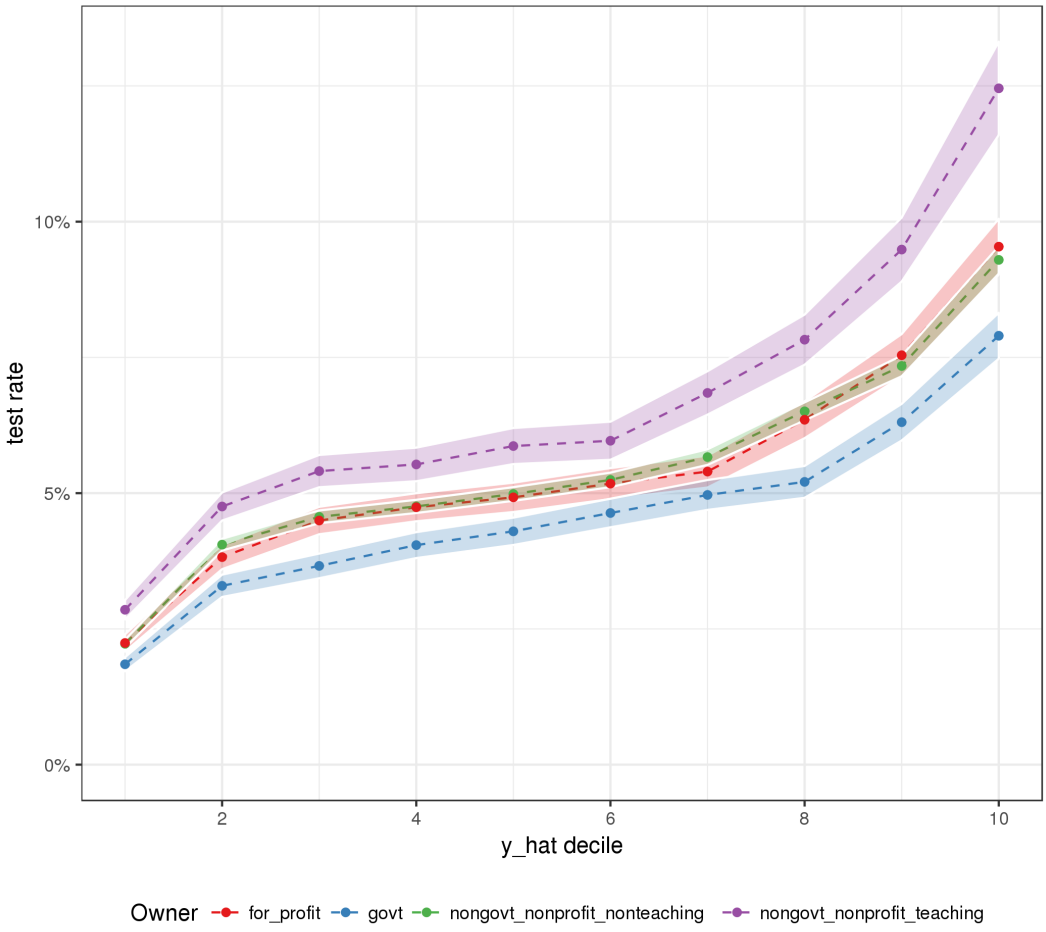


Figure 17: Testing rate as a function of model-predicted risk, by hospital ownership. The groups are defined based on American Hospital Association data, and include teaching hospitals (top: purple), for profit (middle: red), federal hospitals (bottom: blue), and all others, largely non-profit (middle: green).

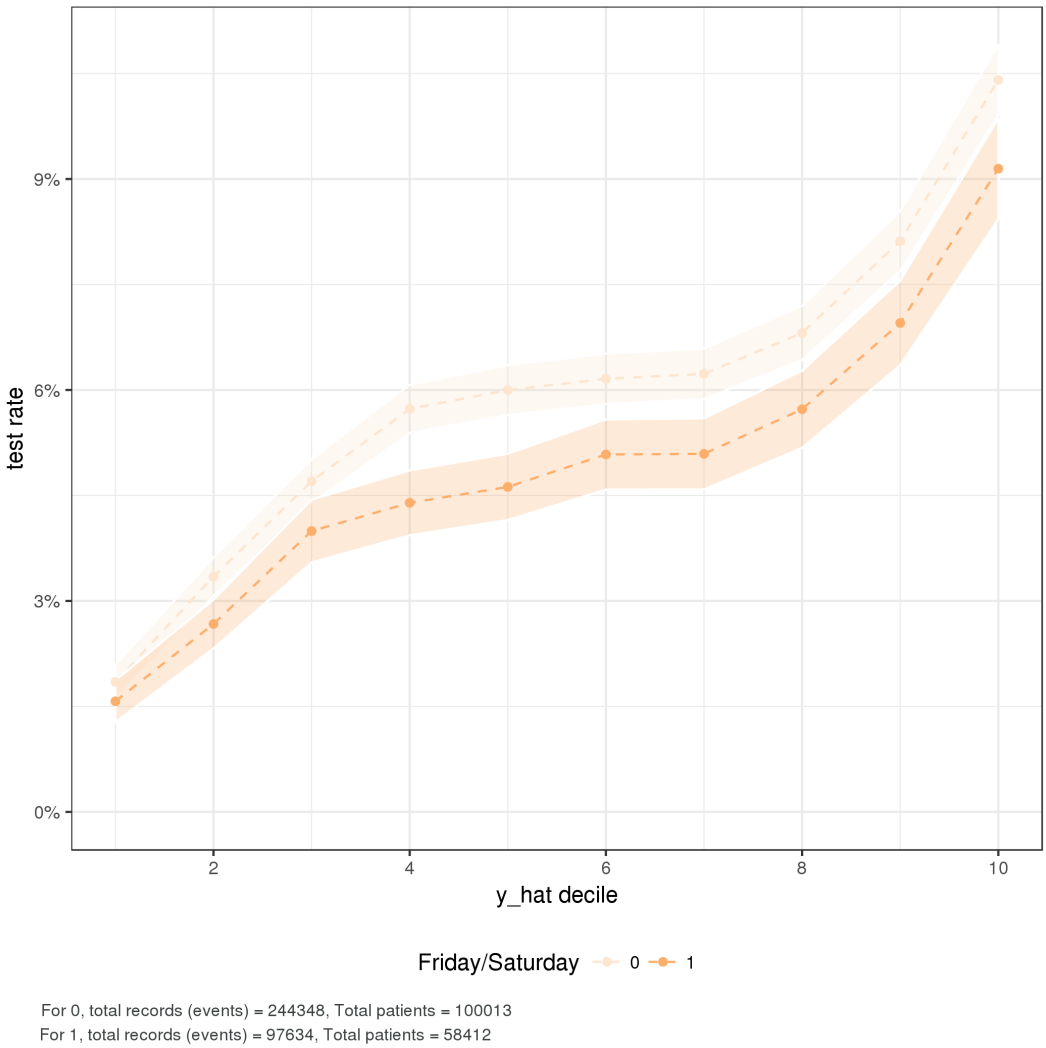


Figure 18: Testing rate as a function of model-predicted risk, by whether the ED visits occurred on Sunday through Thursday (i.e., visits for which stress tests and catheterizations are easily available the following day: top) or Friday and Saturday (for which testing the next day is impossible or more difficult: bottom).

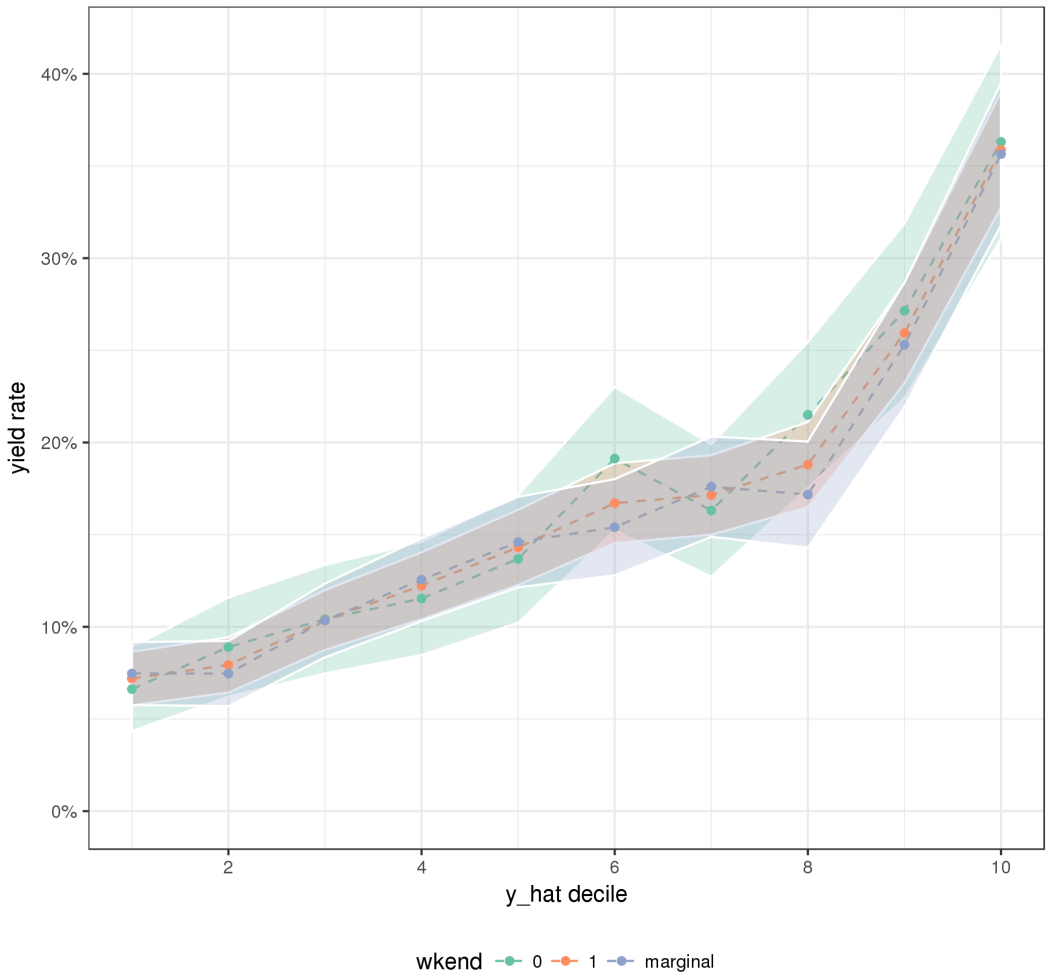


Figure 19: Yield of testing as a function of model-predicted risk, for (1) visits occurring on Sunday through Thursday; (2) visits occurring on Friday and Saturday; and (3) calculated marginal yield for those tested on Sunday through Thursday, but not Friday and Saturday.

Table 1: Top 20 factors predicting over- and under-testing conditional on risk

Over-testing		**Under-testing**	
Variable	*Estimate*	*Variable*	*Estimate*
Chief complaints			
Cardiac arrest	0.618	Headache or pressure	-0.068
Acute coronary syndrome	0.508	Back complaint	-0.055
Congestive heart failure	0.219	Arm or leg weakness	-0.053
ECG abnormality	0.185	Blood in urine	-0.035
Chest pain	0.134	Low blood sugar	-0.034
Knee complaint	0.051	Abdominal complaint	-0.026
Shortness of breath	0.049	Cellulitis or rash	-0.025
Palpitations	0.044	Stroke	-0.025
Unresponsiveness	0.030	Laboratory abnormality	-0.025
Medical device problem	0.023	Leg swelling	-0.019
Loss of consciousness	0.022	Fall	-0.019
Sickle cell disease	0.021	Coughing up blood	-0.019
Jaw, mouth, or lip complaint	0.015	Suture removal	-0.017
High blood pressure	0.015		
Demographics			
White, Male, 70+	0.024	Other race, Female, <50	-0.027
Black, Male, 70+	0.015	Hispanic, Female, <50	-0.025
		White, Female, <50	-0.025
		Black, Female, <50	-0.019
		Hispanic, Male, <50	-0.017
		Other race, Male, <50	-0.017
Other			
Cardiovascular visits (1 month)	0.033	Number of PET scans (3 years)	-0.017
Stents (3 years)	0.023		
Prior ECG finding: J-point	0.019		
Framingham risk factors	0.015		

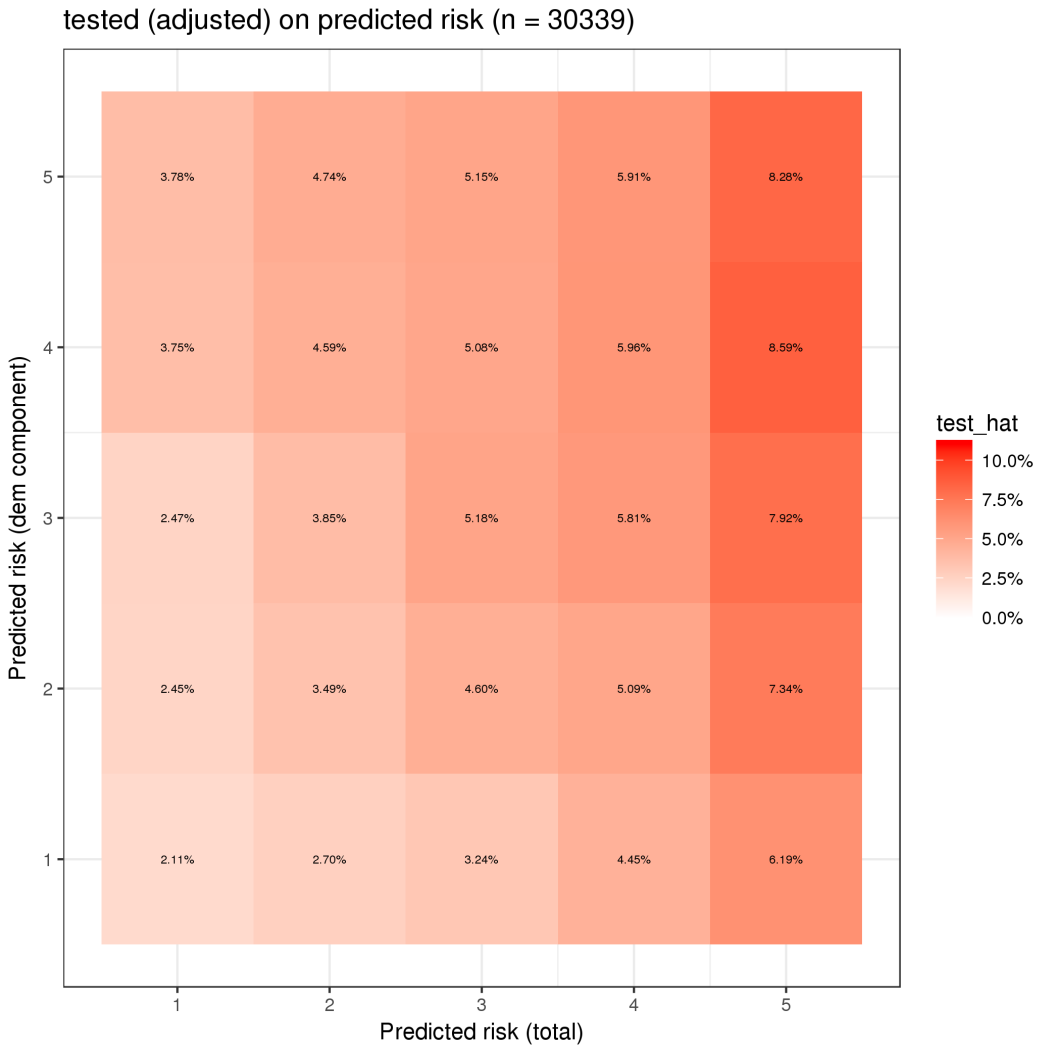


Figure 20: Testing rate as a function of quintiles of model-predicted risk, $\hat{y}$ (horizontal), and $\hat{y}_{dem}$ (vertical: a linear projection of $\hat{y}$ onto demographic groups).

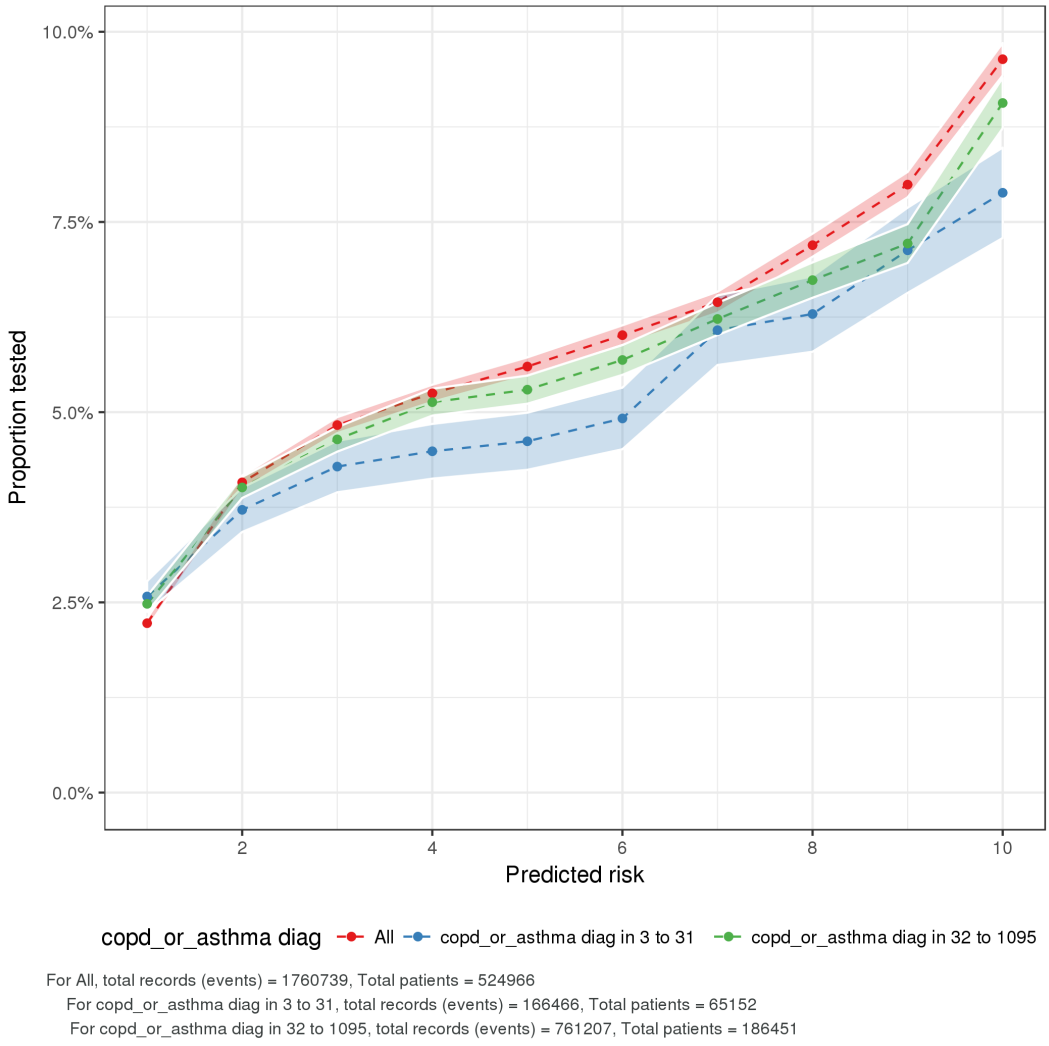


Figure 21: Testing rate as a function of model-predicted risk, for (1) all patients: top (red), (2) patients with a prior diagnosis of COPD or asthma over the three years before ED visits: middle (green), and (3) patients with an encounter for a diagnosis of COPD or asthma in the 30 days before ED visits: bottom (blue).

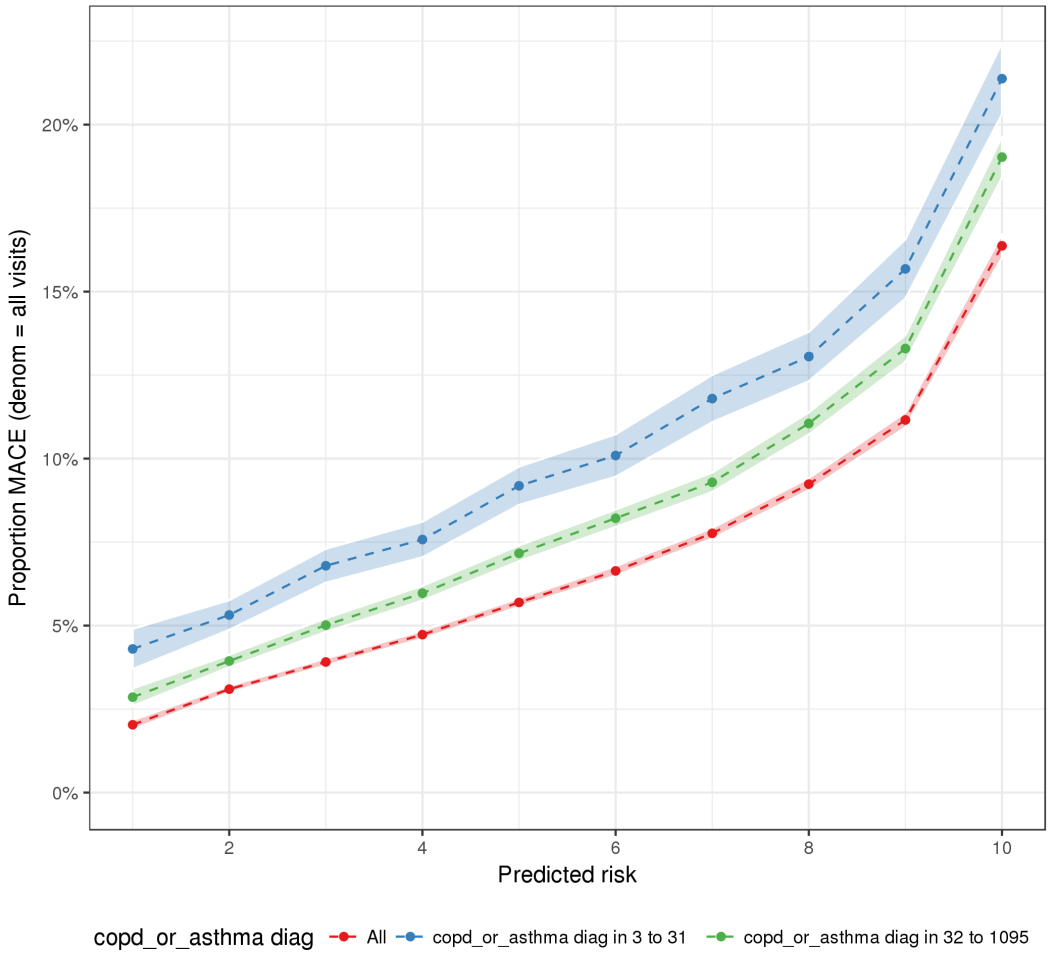


Figure 22: Major advserse cardiac event rate as a function of model-predicted risk, for (1) all patients: bottom (red), (2) patients with a prior diagnosis of COPD or asthma over the three years before ED visits: middle (green), and (3) patients with an encounter for a diagnosis of COPD or asthma in the 30 days before ED visits: top (blue).

Table 2: Comparing Machine Learning to Logistic Regression

Predicted Risk Percentile	ML/Logit Overlap	Average Observed Intervention Rate for Visits Identified as High Risk by:				
		Both ML & Logit	ML Only	Logit Only	All ML Cases	All Logit Cases
1%	26.1%	.4067	.5110	.2767	.5046	.3307
		(.0118)	(.0185)	(.0164)	(.0159)	(.0149)
5%	40.8%	.3344	.3851	.2324	.3961	.3048
		(.0053)	(.0090)	(.0078)	(.0070)	(.0065)
10%	47.6%	.3043	.3341	.2201	.3494	.2889
		(.0037)	(.0066)	(.0057)	(.0048)	(.0045)
25%	63.1%	.2630	.2771	.1762	.2961	.2579
		(.0024)	(.0047)	(.0039)	(.0029)	(.0028)