2016

The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care

Jason Abaluck
Yale University

Leila Agha
Boston University

Chris Kabrhel
Harvard University

Ali Raja
Harvard University

Arjun Venkatesh
Yale University

HHS Public Access

Author manuscript

Am Econ Rev. Author manuscript; available in PMC 2017 November 01.

Published in final edited form as:

Am Econ Rev. 2016 December ; 106(12): 3730–3764. doi:10.1257/aer.20140260.

The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care*

Jason Abaluck,
Yale University and NBER

Leila Agha,
Boston University and NBER

Chris Kabrhel,
Massachusetts General Hospital and Harvard University

Ali Raja, and
Brigham & Women's Hospital and Harvard University

Arjun Venkatesh
Yale-New Haven Hospital and Yale University

Abstract

A large body of research has investigated whether physicians overuse care. There is less evidence on whether, for a fixed level of spending, doctors allocate resources to patients with the highest expected returns. We assess both sources of inefficiency exploiting variation in rates of negative imaging tests for pulmonary embolism. We document enormous across-doctor heterogeneity in testing conditional on patient population, which explains the negative relationship between physicians' testing rates and test yields. Furthermore, doctors do not target testing to the highest risk patients, reducing test yields by one third. Our calibration suggests misallocation is more costly than overuse.

1 Introduction

Many have argued that current medical practice involves large amounts of wasteful spending, with little cross-sectional correlation between regional health spending and health outcomes (Wennberg et al. 1996). But determining the best approach to lower costs and improve quality depends critically on the nature of the inefficiency (Garber and Skinner 2008): is the problem that physicians are spending to the "flat of the curve" where marginal

*An earlier draft of this paper circulated under the title, "Negative Tests and the Efficiency of Medical Care: What Determines Heterogeneity in Imaging Behavior?" Thanks to Brian Abaluck, Joe Altonji, Joshua Aronson, David Chan, Judy Chevalier, Michael Dickstein, David Dranove, Amy Finkelstein, Howard Forman, Jonathan Gruber, Nathan Hendren, Vivian Ho, Mitch Hoffman, Lisa Kahn, Jon Kolstad, Amanda Kowalski, Danielle Li, Costas Meghir, David Molitor, Fiona Scott-Morton, Blair Parry, Michael Powell, Constana Esteves-Sorenson, Ashley Swanson, Bob Town, and Heidi Williams as well as seminar participants at AHEC 2012, AEA meeting 2013, Boston University, Cornell, HEC Montreal, IHEA 2013, the National Bureau of Economic Research, NIA Dartmouth research meeting, the National Tax Association annual meeting, Northwestern, Stanford, University of Houston, and Yale. Funding for this work was provided by NIA Grant Number T32-AG0000186 to the NBER.

returns to treatment are low, or are physicians treating the wrong patients and achieving suboptimally low returns for a given amount of spending?

Diagnostic imaging has been a particularly salient target for policy intervention to prevent overuse. Use of imaging studies grew faster than any other physician service between 2000 and 2007 (Iglehart 2009), leading to concerns about the costs and appropriateness of these imaging tests (Rao and Levin 2012). The Choosing Wisely campaign, sponsored by the American Board of Internal Medicine Foundation and other leading professional societies, encouraged reductions in use of 45 common tests and procedures in 2012, over half of which were diagnostic imaging services.

In this paper, we develop an econometric framework for evaluating how testing intensity and selection of patients impact yields of diagnostic imaging studies. To identify testing intensity (defined as the tendency to test any given patient), our framework decomposes variation in diagnostic imaging rates across doctors into heterogeneity in patients' benefits from testing and heterogeneity in physicians' tendency to test a given patient. Additionally, the model identifies whether physicians are weighting patient observable risk factors to maximize test yield (i.e. the number of positive tests for a given number of tests). Despite the widespread policy attention to the problem of overuse in imaging, our analysis finds that the welfare costs of misallocation are much larger than the costs of overuse. Our findings suggest that for a popular and common diagnostic test, physicians systematically fail to target imaging to those patients with the greatest risk of an acute, often fatal medical condition.

Our model builds on classical econometric selection models originally developed by Heckman (1979) and refined by Chandra and Staiger (2011). Adapting these models to study repeated test decisions by physicians, we argue that the test yield among each doctors' marginally tested patients—those tested patients whom the doctor is nearly indifferent between testing and not testing—can be used to reveal the doctor's testing intensity and provides exclusion restrictions useful for identifying whether doctors successfully maximize test yields.

The same modeling approach can be applied in any setting where we observe repeated choices by a decision-maker meeting two conditions: first, the decision-maker aims to maximize an observable outcome among selected individuals; second, the value of the relevant outcome is known under the counterfactual where selected individuals were not selected.[1] In this case, we assume that physicians seek to maximize test yield for a given number of tests, and we know that test yield is zero if a patient is not tested (the condition will not be detected without this test). Other applications include banks deciding which customers to loan to at a given interest rate in order to maximize profits or employers deciding which employees to hire to maximize productivity. Banks earn zero profits from

[1]We discuss the second condition at greater length in Section 4. In Abaluck, Agha, and Chan (2016), we extend the framework developed here to the case of two-sided selection also studied by Chandra and Staiger (2011). Specifically, we study the decision of whether to treat a patient with Warfarin to minimize strokes; unlike the case studied in this paper where knowing test yield fully reveals the impact of testing on the probability of a positive test (and given calibration assumptions, on the medical value of the test), knowing strokes only among treated patients does not suffice to recover treatment effects for those patients.

customers who do not receive a loan and employers get no productivity benefits from employees who are not hired, so our second condition is satisfied.

We apply our model to analyze CT scans that test for pulmonary embolism (PE). Estimation of the model requires that we can observe test outcomes among patients selected for testing, as well as the structural assumption that doctors will order a CT scan to test for PE if the patient's ex ante risk of PE exceeds a doctor-specific testing threshold. This threshold is our patient invariant measure of physician testing intensity and we seek to recover it for each doctor in our sample.

Identifying differences in physicians' practice styles separately from patient heterogeneity typically requires either quasi-random assignment of patients to physicians or estimates of potentially heterogeneous causal effects of medical treatment for each patient. Prior research, including Chandra and Staiger (2011) and Currie and MacLeod (2013), has argued that reliable estimates of causal treatment effects can be obtained using detailed chart data to control for all patient characteristics observable to doctors, but such data is typically only available in limited samples. This stumbling block makes it difficult to investigate both the extent and the determinants of healthcare overuse or misuse.

A key insight of this paper is that the *ex post* value of a diagnostic test, in this case chest CT scans, is partially observable in insurance claims records based on whether the test results in the relevant diagnosis. A doctor who performs many negative CT scans, which have little *ex post* value for improving patient health, is likely to have a low testing threshold. Our model accounts for heterogeneity in patient PE risk and shows how to recover physicians' testing thresholds. Using these estimated testing thresholds, we investigate the role of medical training, malpractice environment, hospital characteristics and regional factors in shaping practice styles. The model also allows investigation of whether doctors are misweighting observable patient risk factors in selecting which patients to test for PE. By comparing how observable risk factors predict physicians' testing decisions to how those same variables predict rates of positive tests amongst tested patients, we can identify whether physicians are targeting CT scans to the patients with the highest risk of PE based on demographics and comorbid conditions.

Previous research has identified important differences in practice style and skill across physicians. Chandra and Staiger (2011) conclude that overuse of care explains a large amount of variation in treatment for heart attacks across hospitals. Currie and MacLeod (2013) uncover substantial heterogeneity in diagnostic skill across obstetricians. Finkelstein et al. (2014) find that roughly half of the variation in medical spending across regions is driven by provider behavior (rather than patient preferences or health risks), and Molitor (2012) reports that environmental factors explain much of the variation in physician's rates of cardiac catheterization.

We extend this prior literature by not only estimating heterogeneity in physician practice styles, but also explicitly demonstrating that differences in practice style explain why physicians who use more medical resources have lower average medical returns to utilization. We then estimate the resulting welfare loss from the measured variation in

practice styles. We additionally investigate physicians' systematic underweighting and overweighting of patient risk factors and assess how failure to target medical resources to the patients with the highest expected returns impacts health benefits and total welfare. To our knowledge, we are the first to do so in the health economics literature. This analysis highlights a policy-relevant mechanism by which physician decisions may influence health outcomes, and sheds light on the economic importance of these systematic errors in expert judgment.

We analyze 1.9 million emergency department visits drawn from a 20% sample of Medicare claims data, 2000–2009. We present reduced form evidence of a sharply negative relationship between physician testing rates and test yields: those physicians who test the most patients also have the lowest rate of positive tests. We apply a structural model to show that this pattern is explained by enormous heterogeneity in doctors' testing thresholds. Doctors who test more patients move further down the net benefit curve and test patients who are less likely to test positive. Less experienced doctors and doctors in higher spending regions tend to have lower risk thresholds at which they deem CT imaging worthwhile.

Further, physicians fail to target the test to the highest risk patients. Recognized risk factors based on a patient's medical history, some of which are included in popular PE risk scores, continue to receive too little weight in physicians' testing decisions. On the other hand, symptoms appear to be overweighted in some cases. Physicians tend to overtest patients previously diagnosed with one of several conditions which have similar clinical symptoms to PE: rather than infer the patient is having a recurrent episode of their existing condition, the physician may order a PE CT despite the low predicted risk. Finally, black patients are tested less often than other patients despite their higher risk of PE.

Applying calibration assumptions about the cost of testing, the benefits of treating PE and the likelihood of false positives, we compare our estimated distribution of physician testing thresholds to the calibrated socially optimal threshold. This comparison tells us whether doctors are overtesting or undertesting from a social standpoint.[2] Under our preferred calibration assumptions, 84% of doctors are overtesting in the sense that for their marginal tested patients,[3] the costs of testing exceed the benefits. In a simulation where no doctors overtested, the net social benefits from chest CTs would increase by 60% and the number of chest CT scans would fall by 50%. The calibration also allows us to assess the degree of inefficiency from physician misweighting of patient risk factors. Weighting observable comorbidities to maximize test yields would increase the net benefits of testing by more than 300%, primarily by leading to additional testing and appropriate diagnosis of patients with a PE.

[2]Earlier drafts of this paper called this an "allocative inefficiency". In the framework of Garber and Skinner (2008), this is an allocative inefficiency in the sense that one has gone too far along the flat of the curve relating health outcomes to spending, meaning that the marginal return to an additional dollar of care is small (too many resources are allocated to this service). This is contrasted with a productive inefficiency in which one is on a lower production function than could feasibly be achieved. Confusingly, such a "productive inefficiency" may well result from misallocation of resources - for example, failing to allocate CT scans to those patients who benefit most. To avoid the resulting confusion, we now avoid the use of the terms "allocative inefficiency"" and "productive inefficiency" and only use the term "allocation" in the context of whether physicians are appropriately choosing which patients to test in order to maximize test yield.

[3]Throughout this paper, "marginal" patients is used to refer to those patients whom a given physician is indifferent between testing and not testing.

The paper is organized as follows. Section 2 provides some background on chest CT scans for PE. Section 3 describes the data and uses reduced form evidence to motivate the structural model. Section 4 lays out our structural model of testing behavior and describes our estimation strategy. Section 5 reports results from estimating our structural model. Section 6 probes the robustness of these results to alternative modeling approaches that relax or vary key identifying assumptions. Section 7 conducts simulations to uncover the welfare implications of our findings, and Section 8 concludes.

2 Background on PE CTs

We study testing behavior in the context of chest CT scans performed in the emergency department to detect PE. PE is the third most common cause of death from cardiovascular disease, behind heart attack and stroke (Goldhaber and Bounameaux 2012), and CT scans are the primary tool for diagnosis of PE. Yet given the financial costs and medical risks of testing, PE CT scans are commonly thought to be overused in emergency care. The American College of Radiology targeted PE CT as a key part of the *Choosing Wisely* campaign aimed to reduce overuse of medical services. Despite the concern about overuse, the Office of the Surgeon General (2008) estimates that approximately half of PE cases are undiagnosed, based on analysis of autopsy reports. The simultaneous concern in the medical community about overuse and missed diagnoses raises the question of whether diagnostic testing for PE is currently being targeted to maximize PE detection.

A PE occurs when a substance, most commonly a blood clot that originates in a vein, travels through the bloodstream into an artery of the lung and blocks blood flow through the lung. It is a serious and relatively common condition, with an estimated 350,000 diagnosed cases of PE per year in the United States (Office of the Surgeon General 2008). Left untreated, the mortality rate from a PE depends on the severity and has been estimated to be 2.5% within three months for a small PE (Lessler et al. 2010), with most of the risk concentrated within the first hours after onset of symptoms (Rahimtoola and Bergin 2005). Accurate diagnosis of PE is necessary for appropriate follow-up treatment; even high risk patients are unlikely to be treated presumptively.

CT scans to test for PE have a number of attractive features for our purposes: they are a frequently performed test; they introduce significant health risks and financial costs; a positive test is almost always followed up with immediate treatment, observable in Medicare claims records; and a negative test provides little information to the physician about alternative diagnoses or potential treatments. We discuss each of these features in more detail in Appendix B, explaining how the clinical context supports our modeling assumptions.

PE is an acute event with a sudden onset. The symptoms of PE are both common and nonspecific: shortness of breath, chest pain, or bloody cough. Hence, there is a broad population of patients who may be considered for a PE evaluation. Practice guidelines recommend that physicians also consider several additional risk factors before determining whether to pursue a workup for PE.[4]

Many argue that PE CT scans are widely overused (Coco and O'Gurek 2012, Mamlouk et al. 2010 and Costantino et al. 2008). Recent estimates by Venkatesh et al. (2012) suggest that one third of CT scans in a sample of 11 US emergency departments would have been avoidable if physicians had followed National Quality Forum guidelines on CT usage. The nonspecific symptoms of PE and significant mortality risk likely both contribute to overuse, particularly in the emergency care setting.

A CT angiogram is the standard diagnostic tool for PE. The average allowed charge in the Medicare data is around $320 per PE CT when the bill is not covered by a capitation payment. Payment goes to the radiologist for interpreting the scan and to the hospital for the technician and capital equipment required to perform the scan. The emergency department doctor responsible for ordering the test has, at most, a diffuse incentive to ensure the hospital's financial health and reduce his malpractice risk, but he receives no direct payments from Medicare or the hospital for ordering a scan.

PE CT scans also come with small but important medical risks. The most significant risk arises from false positive CT scans which lead to additional unnecessary treatment with anticoagulants, incurring financial costs and creating significant risk of bleeding. In addition, there is an estimated 0.02% chance of a severe reaction to the contrast, which then carries a 10.5% risk of death (Lessler et al. 2010), although this cost is small relative to the billed financial costs of a CT scan. Finally radiation exposure may increase downstream cancer risk, although the additional lifetime cancer risk is minimal for the elderly Medicare population in this study.

The key simplifying assumption we make to evaluate the net benefits of testing is that a negative test has no value. This assumption is not true in general for all tests: a negative test may rule out one treatment thus justifying treatment for an alternative, or a negative test might prevent an otherwise costly treatment. However, in our setting—CT scans for PE—a positive test is followed by an inpatient admission and treatment with blood thinners while a negative test does not suggest any further interventions or testing for related problems. We defend this assumption at greater length in Appendix B.

3 Data

We combine data from five sources: Medicare claims records, the American Hospital Association annual survey, the American Medical Association Masterfile, the Medicare Physician Identification and the Eligibility Registry, and the Avraham Database of State Tort Law Reforms. Using a 20% sample of Medicare Part B claims from 2000 through 2009, we identify patients evaluated in an emergency department and observe whether they were tested for PE, as well as whether any such test succeeded in detecting PE.

[4]Popular practice guidelines use the following factors to calculate a risk score: age, elevated heart rate, recent immobilization or surgery, history of deep vein thrombosis or PE, recent treatment for cancer, coughing up blood, lower limb pain or swelling, and chances of an alternative diagnosis.

3.1 Medicare claims data

We begin by identifying all patients evaluated in the emergency department (ED), using physician-submitted Medicare Part B claims for evaluation and management.[5] The physician submitting this claim for evaluation and management is responsible for the patient's emergency care; it is his decision whether or not to order testing for PE. Using physician identifiers, we track the behavior of all doctors who routinely evaluate Medicare patients in the ED.

We identify which ED patients are tested for a PE using bills submitted by radiologists for the interpretation of chest CTs with contrast, when the CT is performed within 1 day of the ED visit.[6] We restrict our sample to physicians who order at least seven in sample CT scans between 2000–2009, since very low volume doctors provide too little information to accurately estimate physicians' testing thresholds.[7]

While diagnosis of PE is the most common purpose of a chest CT performed in the emergency care setting, there are a small handful of other, less common indications, including pleural effusion, chest and lung cancers, traumas, and aortic dissection. For this reason, we exclude patients from the sample who are coded with a diagnosis related to trauma, pleural effusion, chest or lung cancer, or patients with a history of aortic aneurysm, aortic dissection, or other arterial dissection. We also exclude patients with a history of renal failure, since these patients are likely ineligible for a CT scan with contrast, due to risks of the contrast agent. These sample restrictions are designed to limit the sample to patients who may be eligible for a chest CT scan and for whom the scan is highly likely to have been ordered to detect PE; these assumptions are discussed in more detail in Appendix C.

Once we have identified relevant CT scans in billing data, we then need to code the test outcome, i.e. whether or not the scan detected a PE. Patients with acute PE are typically admitted to the hospital for monitoring and to begin a course of blood thinners or place a venous filter to reduce clotting risk. From the sample of patients tested in the emergency department with a chest CT, we identify positive tests on the basis of Medicare Part A hospital claims that include a diagnosis code for PE among any of the diagnoses associated with the hospital stay.

We have validated this approach to identifying positive tests by using cross-referenced patient chart and hospital billing data from two large academic medical centers. The evidence from these centers suggests that we are unlikely to understate physicians' testing thresholds due to undercounting of positive test results. More detail on this data validation exercise is presented in Appendix D.

In addition to measuring whether patients were tested and the testing outcome, we also document a number of characteristics that allow us to predict the patient's propensity to be

[5]In particular, we identify patients based on CPT codes for emergency department evaluation and management: 99281, 99282, 99283, 99284, 99285, and place of service 23 (i.e. hospital emergency department).
[6]We begin by identifying all bills for chest CTs on the basis of CPT codes 71260, 71270, and 71275.
[7]In our sample, this restriction drops about 1/2 of all CT scans since a large number of patients are evaluated by very low volume providers. Nonetheless, our sample likely includes the most policy relevant sample - it is difficult to target interventions at physicians who order a procedure less than once a year.

diagnosed with a PE, including age, race, sex, and medical comorbidities. We code comorbidities from both Medicare's Chronic Condition Warehouse and from the Elixhauser et al. (1998) definitions; while these sets of conditions overlap, the Chronic Condition Warehouse utilizes outpatient claims to code comorbidities whereas the Elixhauser comorbidities are based only on inpatient medical history, so they typically encode different levels of disease severity. We augment these standard sets of medical comorbidities to include several measures that are specific to PE risk: whether the patient was previously admitted to the hospital with a diagnosis of PE, thoracic aortic dissection, abdominal aortic dissection, or deep vein thrombosis, and any cause admission to the hospital or surgical hospital admission within 7 days or 30 days.

3.2 Physician, hospital, and regional data

After using the Medicare claims data to estimate the testing threshold applied by each doctor, we explore predictors of physicians' practice styles by linking testing thresholds to physician, hospital, and regional characteristics.

We draw physician data from two sources, the Medicare Physician Identification and Eligibility Registry (MPIER) and the American Medical Association Masterfile (AMA data). The MPIER and AMA both identify the medical school and graduation year for each physician, which we have linked to the US News & World Report medical school rankings. We bin schools according to whether they are typically ranked in the top 50 for either primary care or research rankings.

Hospital characteristics are drawn from the American Hospital Association annual survey. We use these data to observe whether the physician typically practices at a for profit hospital or an academic hospital, defined as a hospital with a board certified residency program.

Using provider zip codes, we identify the hospital referral region (HRR) in which each patient is treated. HRRs are regional health care markets defined by the Dartmouth Atlas to reflect areas within which patients commonly travel to receive tertiary care. There are 306 HRRs in total. Using data from the Dartmouth Atlas, we link each HRR to the average spending per Medicare beneficiary to capture a broad measure of regional care intensity.

Finally, data on state malpractice environment is from Avraham (2011) Database of State Tort Law Reforms. Following prior work by Currie and MacLeod (2006) and Avraham et al. (2012), we focus on two key measures of malpractice law: whether a state has enacted malpractice damage caps on award amounts, and joint and several liability reform.

3.3 Summary statistics

There are 1.9 million emergency department visit evaluations in our dataset, after making the sample exclusions noted above. Of these patients evaluated in the ED, 3.8% of them are tested with a chest CT scan with contrast. Amongst tested patients, 6.9% of them receive a positive test, i.e. are admitted to the hospital within 24 hours with a diagnosis of PE.

Summary statistics are reported in Table 1, with results reported separately for patients who do not receive a CT scan (column A), patients who receive a negative test (column B), and

patients with a positive test (column C). We observe the testing behavior of over 6600 physicians, with an average of 284 ED patients per physician.

Patient demographics are similar across the untested and tested patient groups. The average age is 78 years in the untested sample and slightly lower (77 years) in the sample of patients with negative or positive tests. Patients who test negative are more than twice as likely to have a history of PE as untested patients; patients with positive tests are five times more likely to have a history of PE than untested patients.

We note a few modest differences in physician background and practice environment across patient groups. Patients with negative tests are evaluated by doctors with five months less experience on average than patients with positive tests, and were treated in regions with 1% higher Medicare spending per beneficiary, compared to patients with positive tests. Among tested patients, those with positive tests were 1 percentage point more likely to have been evaluated by a doctor trained at a top tier medical school. In the structural model, we will decompose to what extent these differences may be driven by differential sorting of high risk patients and to what extent they reflect differences in physician practice styles.

3.4 Reduced form evidence of heterogeneity in doctor testing behavior

Before describing our model, we consider reduced form evidence of heterogeneity in doctors' testing behavior. We first divide doctors in our sample into 10 deciles according to the average fraction of patients tested. We observe average testing rates that range from 1.7% of ED patients in the lowest physician decile to 8.2% of ED patients in the highest physician decile. We want to know whether this variation reflects differences in doctor behavior for patients with similar PE risk, or differences in patient PE risk for physicians with similar testing intensities.

We can separate these hypotheses by comparing rates of positive tests conditional on testing behavior. If doctors who test more do so because their patients are at higher risk of PE, we should expect that doctors with higher testing rates will also have a higher fraction of positive tests among tested patients.[8] Alternatively, if doctors who test more do so because they are the type that tests more for any given level of patient risk, then we expect to find that physicians who test more also have a lower fraction of positive tests among tested patients. In the latter case, physicians could differ in the threshold probability at which they think testing is worthwhile, and physicians who test more are moving further down the expected benefits curve.

To illustrate this point, we have sketched a stylized picture of the testing decision in Figure 1. Patients are sorted along the x-axis according to their risk of PE, q_{id}, from highest risk to lowest risk. The x-axis corresponds to the cumulative fraction of patients, and the y-axis corresponds to the marginal patient's PE risk q_{id}, so that each point (x, y) along the plotted curve shows the fraction x of patients for whom $q_{id} \geq y$. For example, at point ($T^A = 2/3$, τ^A

[8] In particular, both doctors would have similar test yields among marginal tested patients, but the doctor who tests more would have a higher test yield among the higher risk inframarginal patients. We formalize the points in this section in the context of our structural model in section 4.

= 1/2) in Panel A, the graph indicates that 2/3 of patients have a risk of PE that equals or exceeds 1/2. (We use this unrealistically high risk for illustrative purposes.)

In Panel A, we consider two doctors with the same patient distribution of PE risk, but with different testing thresholds. Doctor A tests every patient whose personal PE risk q_{id} exceeds Doctor A's testing threshold τ^A, and likewise Doctor B tests all patients for whom $q_{id} > \tau^B$. Because Doctor B's threshold is lower than Doctor A's, i.e. $\tau_B < \tau_A$, Doctor B tests a greater fraction of patients, $T^B > T^A$. Doctor B's tested patients have a lower average PE risk than Doctor A's tested patients, so Doctor B's test yield Z^B—i.e. the fraction of positive tests among tested patients—is lower than Doctor A's test yield Z^A, as can be seen in the graph. In this panel, there is a downward sloping relationship between the fraction of patients each doctor tests and his average test yield.

In Panel B, we consider an alternate scenario which could also explain why Doctor B continues to test a greater fraction of his patients than Doctor A, i.e. why $T^B > T^A$. In this example, doctor A and Doctor B have the same testing threshold, so $\tau_B' + \tau_A'$. Given the same expected patient PE risk, Doctors A and B would arrive at the same testing decision. However, the two doctors now face different distributions of patient PE risk. For any given probability of a positive test, Doctor B sees (weakly) more patients with q_{id} exceeding the common threshold for testing. In other words, Doctor B's patient population is higher risk than Doctor A's. As can be seen in the graph, Doctor B's test yield $Z^{B'}$ will be higher than Doctor A's test yield $Z^{A'}$, even though both doctors have the same testing threshold, since more of the mass in Doctor B's distribution of patient risk is concentrated at higher risk levels. In contrast with Panel A, there is now an upward sloping relationship between the fraction of patients each doctor has tested and his average test yield.

Now turning to our observed Medicare data, we use a simple binned scatterplot to explore whether variation in risk for PE or variation in testing behavior can explain the differences in physicians' testing propensities. We begin by binning physicians into deciles according to the fraction of patients they test; next we calculate the fraction of tested patients for whom PE was detected within each decile. This relationship between fraction tested and average test yield is plotted in Figure 2 Panel A. The graph displays a generally downward sloping relationship between average testing probability along the x-axis and fraction of tested patients with detected PE along the y-axis. Doctors who test a greater fraction of their patients are less likely to find positive test outcomes among tested patients; a simple regression reveals this relationship is highly significant. The figure suggests that differences in testing thresholds across doctors may be an important determinant of observed heterogeneity in testing behavior. It appears that doctors who are more likely to test their patients compared to their peers are also testing more low-risk patients.

Our structural model formalizes the intuition described above. It is designed to disentangle (observable and unobservable) differences in patient PE risk from differences in physician testing thresholds and evaluate the contribution of each to observed variation in testing behavior, following the intuition of this simple empirical exercise. We discuss the structural model in more detail in Section 4 below.

3.5 Reduced form evidence of misweighting patient PE risk factors

In addition to considering heterogeneity in physicians' testing thresholds, we also investigate whether physicians are successfully identifying observable risk factors associated with the highest probability of positive tests and testing patients with those characteristics. Determining which patients should be tested requires complex, subtle judgments about clinical risk on the basis of many factors. In our data, we capture some of the most common and relevant comorbidities by analyzing patients' claims histories. Guided by the structural analysis that follows, we motivate our exploration of misweighting PE risk with a few simple examples.

Consider a comparison of patients with a history of prostate cancer to those with no such history. Patients with a history of prostate cancer are no more likely to be tested for PE than patients without that condition; in fact, testing rates are slightly lower among prostate cancer patients (3.7%) compared to the rest of the population (3.8%). However, it turns out that among tested individuals, prostate cancer patients are over 50% more likely to be diagnosed with PE than patients with no such history.

In Figure 2 Panel B, we see that for each decile of doctors' overall testing rate, doctors are equally or more likely to test patients without prostate cancer, despite the consistently higher PE risk among patients with prostate cancer. As described in the previous section, in the absence of variation in physician practice style, we would expect this graph to be upward sloping: doctors who tested more patients would do so because they have higher risk patients and higher expected test yields. Splitting the sample by comorbidity, if patients with a given comorbidity have higher yield they should also be tested at higher rates.

A PE risk score popularly used to guide physicians on whether to order diagnostic testing includes treatment for cancer malignancy among its 7 risk criteria (Wells et al. 1995; Wells et al. 1998; Wells et al. 2000). And yet, although cancer is a recognized clinical risk factor for PE, a relationship supported by our data, it appears that patients with a history of prostate cancer are no more likely to be tested than the average ED patient. This provides the first suggestive evidence that physicians may not be properly accounting for the increased PE risk associated with prostate cancer, and thus may be under-testing prostate cancer patients relative to the rest of the population.

In Table 2, we highlight the basic summary statistics for eight of the clinical factors that show significant evidence of misweighting in the structural model that follows. Similar to the case of prostate cancer, we find that black patients are less likely to be tested than non-black patients, even though among tested patients, the rate of positive tests is much higher for black patients. Figure 2 Panel C illustrates the lower test rates and higher test yield of black patients within every decile of physician test rate. A reverse pattern holds for patients with ischemic heart disease, atrial fibrilation or chronic obstructive pulmonary disease (COPD); they are tested at similar or higher rates than patients without those conditions, despite the fact that tested patients with these conditions are approximately 30% *less* likely to have a PE detected. Figure 2 Panel D shows the test rates are substantially higher and yields lower for pateints with COPD, within each decile of physician test rate.

For other conditions, physicians respond in the right direction but overweight or underweight that condition relative to what would maximize the incidence of positive tests. The model implies that, everything else held equal (including other patient characteristics and physician thresholds), two comorbidities which have the same marginal impact on testing behavior should also have the same marginal impact on the conditional likelihood of a positive test. Our model identifies a few factors which appear to have a disproportionate impact on the likelihood of a positive test given their impact on testing behavior: a past history of PE, deep vein thrombosis, or a recent hospital admission are associated with 20 to 90 percent higher rates of testing but are 140 to 200 percent more likely to have a PE detected, a disproportionate increase relative to other factors in our model with a similar impact on testing behavior.

This exploration of misweighting presumes that patients with and without a particular risk factor don't differ in their other comorbidities and are sorting to ED physicians with similar testing thresholds. In the structural model, we formalize this analysis, explicitly modeling differences in testing rates that may be driven by physician's testing thresholds or other PE risk factors.

4 Model of testing behavior

Our reduced form results suggest that physicians vary in their testing intensity and that physicians may not be allocating tests in a way that maximizes test yields. Our structural model embeds both possibilities and allows us to assess the quantitative importance of each inefficiency.

First consider the question of how we can identify variation in physicians' practice style. In a world with random assignment of patients to doctors, a simple comparison of average testing rates across doctors could recover physicians' testing intensities, since there would be no cross-doctor variation in patients' ex ante PE risk. After adjusting for statistical noise, the variation in physician testing rates with random patient assignment would tell us whether physicians vary in their testing intensity for identical patients. Unfortunately, in our setting —as in many cases of interest—patients are not randomly assigned to physicians. If we regressed testing behavior on physician fixed effects, those fixed effects would jointly capture both physicians' testing tendencies and the suitability of each physician's patient population for testing. Since some doctors see patients with greater ex ante risk, we cannot attribute all variation in physician fixed effects to differences in doctor testing intensities.

To recover a measure of practice style that is purged of variation due to different patient populations, we apply an insight from Chandra and Staiger (2011) (hereafter, CS) which builds upon classical selection models developed by Heckman (1979) and Heckman and MaCurdy (1980). CS's insight closely parallels the logic in Section 3.4. In their model, if physician A is more inclined to treat any given patient than physician B, then physician A's marginal patients should have lower returns to treatment. Estimation of the original CS model requires observing the individual-specific return to treatment for all treated individuals, a difficult object to recover if one does not have random patient assignment. We

adapt the model to cover diagnostic testing, where test results (positive or negative) can proxy for the impact of treatment on the treated.[9]

As noted in the introduction, the distinctive feature of positive tests which eliminates the need to separately estimate or assume treatment effects arises in many other settings of interest. The essential ingredient is that we need only observe outcomes among treated individuals; counterfactual outcomes if treated patients had not been treated are known. Test yield is zero among untested patients: PE would not have been detected without this test. As a result, we can use test outcomes among tested patients to analyze whether physicians sucessfully maximize test yields for a given number of tests. Likewise, a bank deciding whether to extend credit learns exactly the profits they received from a given customer once the loan term is reached and they know whether default occurred. Similarly, a firm deciding which applicants to hire knows how much productivity they generated from a given employee once the employee has completed a given employment spell.

In contrast, our model would not directly extend to a doctor deciding which patients to treat with a drug in order to minimize stroke risk. In that case, we would not know the impact of the drug just from observing whether a patient had a stroke, because we would not know what the value of the objective function (strokes) would have been absent treatment. Likewise, if the employer's objective were to maximize the productivity of a given employee regardless of which firm they end up in (perhaps a more reasonable objective for a policy-maker), we would not know whether that employee would have been more productive elsewhere. Our model is not directly applicable in such cases without further structure and assumptions.[10]

In the setting of diagnostic tests, the CS intuition becomes very simple because test yield is 0 among untested patients. Suppose physician A is more inclined to test than physician B, in the sense that physician A tests all patients with a probability of a positive test greater than 4% while physician B tests all patients with a probability of a positive test greater than 5%. Then by looking directly at test yields for each doctor's marginal patients, we can recover her threshold: those patients whom physician A is indifferent between testing and not testing have a test yield of 4%, while those patients whom physician B is indifferent between testing and not testing have a yield of 5%. Our model uses cross-doctor heterogeneity in test yield among marginal patients to identify variation in physician testing thresholds, which tell us whether physicians would behave differently given identical patient populations.

To identify misallocation of tests, we ask whether different patient characteristics which predict the same change in testing probabilities also predict the same change in yield conditional on testing. If we find that, for example, patients with and without prostate cancer are tested at the same rates but patients with prostate cancer are much more likely to test positive conditional on being tested, this suggests that one could increase test yield by testing more patients with prostate cancer.

[9]Given our assumption that negative test results do not improve patient health *ex post*, the testing outcome can proxy for the impact of treatment on the treated, as long as the benefits of treating a detected PE are constant across patients. The clinical basis for this assumption is discussed at greater length in Appendix B.

[10]We extend the model to investigate the use of Warfarin to prevent strokes in Abaluck, Agha and Chan (2016).

For our analysis of both physician practice style and test misallocation, considering counterfactuals requires us to predict how test yield would change if physicians test more or fewer patients with a given set of observable characteristics. Following CS (and more generally Heckman and Vytlacil 2005), we recover this information by estimating the relationship between test yields and testing probabilities (or more precisely, predicted indices of testing propensity). In the exposition below, we make explicit the conditions under which the relationship between test yields and testing probabilities fully captures how marginal benefits decline as physicians test more patients.

Our exposition will proceed as follows. First we lay out the CS model with the adaptation described above—i.e. replacing the returns to treatment with the probability of a positive test—and describe how we can recover each physician's testing intensity. In section 4.2, we extend the CS modeling framework to capture the possibility that physicians may not select patients to test in a way that maximizes test yield. In section 4.3, we discuss how physician thresholds, misweighting, and the degree of selection on unobservables can be jointly identified. In section 4.4, we provide further details on how our model is estimated.

4.1 A Chandra-Staiger Model of Testing

Assume that the suitability of a patient for testing is determined entirely by the *ex ante* likelihood of a positive test. We define q_{id} to be the conditional probability of a positive test for patient i evaluated by doctor d, given all the information available to the doctor:

$$q_{id}=x_{id}\beta+\alpha_d+\eta_{id} \quad (1)$$

where x_{id} are observed patient characteristics (which we assume throughout are normalized to have mean 0 for each doctor), α_d are doctor fixed effects, and η_{id} are factors observable to the doctor but unobservable to the econometrician which impact the likelihood that a test is positive. Note that the inclusion of physician fixed effects and allows the population risk of PE to vary across doctors in ways that are not captured by the included patient covariates.[11]

Following the typical structure of Heckman selection models, we begin by assuming that η_{id} is independently and identically distributed across patients and doctors; we refer to this as the "ignorability assumption" following the prior literature. (We explore relaxing the ignorability assumption in Section 6.) We further assume that η_{id} has full support; note it is also bounded because q_{id} lies between 0 and 1.

Following CS, we make the structural modeling assumption that physicians test if and only if the probability of a positive test q_{id} exceeds a physician-specific threshold τ_d. That is, they test if and only if:

$$Test_{id}=1 \leftrightarrow q_{id}=x_{id}\beta+\alpha_d+\eta_{id}>\tau_d \quad (2)$$

[11] CS interpret α_d to reflect variation in expertise rather than differences in patient population. In our setting, where there is separation between the diagnostician ordering the test and the radiologist conducting it, and less expected skill dispersion in interpreting the test, we focus instead on the possibility that some doctors see a patient population which is ex ante more likely to have PEs.

which implies that:

$$pr(Test_{id}=1)=f(x_{id}\beta+\alpha_d-\tau_d) \quad (3)$$

where the functional form of $f(x_{id}\beta + \alpha_d - \tau_d) = Pr(\eta_{id} > -(x_{id}\beta + \alpha_d - \tau_d))$ depends on the distribution of η_{id}. By estimating equation 3, we can calculate the probability that a patient with a given set of observables is tested by doctor d, which will be a nonlinear function of the testing propensity index $I_{id} = x_{id}\beta + \alpha_d - \tau_d$.

τ_d is our measure of physician treatment intensity holding patient population fixed. Physicians with lower τ_d are more likely to test any given patient: they have a lower threshold probability at which they decide testing is worthwhile. If we had random assignment of patients to physicians, then we would know that $\alpha_d = \alpha$ for all physicians and could recover τ_d directly from estimation of equation 3 (at least up to a normalization constant). Without random assignment, α_d and τ_d are not separately identified from observed testing decisions; to separate them, we will need to use data on test outcomes.

Let Z_{id} denote a binary variable indicating whether the test is positive or negative, which we observe only for tested patients. If every patient were tested, we would observe Z_{id} for the entire sample and could recover β and α_d by estimating the linear probability model implied by equation 1 using OLS. (Of course, if every patient were tested, there would be no variation in doctor testing thresholds.) In practice, we only observe whether a test is positive or negative for those patients whom doctors choose to test, so there is a selection problem; this is the standard selection problem originally studied by Heckman (1979).

Formally, we model testing outcomes as follows:

$$E(q_{id}|Test_{id}=1)=E(Z_{id}|q_{id}>\tau_d)=x_{id}\beta+\alpha_d+E(\eta_{id}|q_{id}>\tau_d)=x_{id}\beta+\alpha_d+h(x_{id}\beta+\alpha_d-\tau_d)=\tau_d+\lambda(I_{id}) \quad (4)$$

where $h(x_{id}\beta + \alpha_d - \tau_d) \equiv E(\eta_{id}|q_{id} > \tau_d) = E(\eta_{id}|\eta_{id} > -I_{id})$ and $\lambda(I_{id}) \equiv I_{id} + h(I_{id})$. Test yields are a function of physician thresholds and the propensity to test.

For marginal patients whom doctors are indifferent between testing and not testing, $\lambda(I_{id}) = E(I_{id} + \eta_{id}|I_{id} + \eta_{id} = 0) = 0$, so $E(q_{id}|Test_{id}) = \tau_d$. If a physician tests all patients with a probability of a positive test greater than 3%, then for marginal patients (with the minimum observed value of I_{id} among tested patients), the positive test probability is exactly 3%. The probablity of a positive test will generally rise among inframarginal tested patients, who are more likely to be tested based on observables and doctor fixed effects than marginal patients.

The binned scatterplot of testing rates and test yields described in section 3.4 can provide some intuition for understanding this model. Variation in testing propensities I_{id} could be driven by differences in patient PE risk, either through differences in observed comorbidities

x_{id} or unobserved population risk α_d. Alternatively, differences in testing propensities could be explained by differences in physician testing thresholds τ_d.

If all variation across doctors in testing behavior were driven by patient PE risk, then physicians with higher average testing propensities will have higher test yields. This relationship is apparent in the last line of equation 4; if we hold τ_d fixed and increase I_{id}, $E(q_{id}|Test_{id} = 1)$ will increase.[12] On the other hand, variation in physician testing thresholds τ_d will lead to a downward sloping relationship between testing propensities I_{id} and test yields $E(Z_{id}|q_{id} > \tau_d)$. This relationship is apparent from the first line of equation 4; if we hold α_d fixed and raise testing propensities by decreasing τ_d, then $E(q_{id}|Test_{id} = 1)$ will decrease. The model derivation formalizes the intuitive argument made in section 3.4, which interpreted the observed downward sloping relationship between doctors' average fraction of patients tested and test yield as evidence of variation in testing thresholds.

In sum, average test yields for marginal patients will reveal testing thresholds τ_d among doctors who evaluate enough marginal patients in our sample. Estimating the relationship between higher testing propensities and higher test yields for physicians with known τ_d will identify the function $\lambda(\cdot)$, which allows us to recover τ_d even for lower volume doctors who do not test marginal patients in our sample. Identification is discussed more formally in Section 4.3. The model so far allows us to identify differences in physician testing intensity for fixed patient populations and to simulate how physician testing behavior and outcomes would change if physician practice styles were more uniform.

4.2 Misweighting of patient risk

A key difference between our model and Chandra and Staiger (2011) is that we extend the model laid out above to allow for the possibility that doctors may not successfully select patients on the basis of observable comorbidities to maximize test yields for a given number of tests. We previously assumed that the coefficients β attached to patient observables when doctors decide which patients to test reflect the true relationship between those characteristics and the likelihood of a positive test. This need not be the case. Doctors may under- or over-weight the importance of different risk factors, so that testing is not necessarily targeted at the highest risk patients.

Assume that each doctor's belief about the probability of a positive test is given by:

$$q'_{id} = x_{id}\beta' + \alpha'_d + \eta_{id} \quad (5)$$

while the actual probability remains:

$$q_{id} = x_{id}\beta + \alpha_d + \eta_{id} \quad (6)$$

[12] This is satisfied as long as $\lambda(I_{id}) = E(\eta_{id} + I_{id}|\eta_{id} + I_{id} > 0)$ is upward sloping in the function I_{id}. This restriction holds for many general distributions of η_{id}, including, for example, under distributions meeting the restriction that η_{id} is symmetric and mean 0.

In this model, doctors test if $q'_{id}>\tau_d$. Note that if $\alpha'_{id} \neq \alpha_d$, $q'_{id}>\tau_d$ can be rewritten as:

$$x_{id}\beta'+\alpha_d+\eta_{id}>\tau_d+\alpha_d-\alpha'_d \quad (7)$$

Thus, it is without loss of generality to assume that $\alpha'_d=\alpha_d$ while noting that one reason for variation in thresholds τ_d is that physicians may have mistaken beliefs about patient PE risk α_d. We cannot distinguish between the case where some physicians test more because they have a lower threshold and the case where some physicians test more because they mistakenly believe their patients are more likely to test positive than is actually the case.[13]

We define the new testing propensity $I'_{id}=x_{id}\beta'+\alpha_d-\tau_d$ to reflect the observed propensity given physician beliefs about β'. With this change, we can rewrite the test outcomes equation:

$$E(Z_{id}|Test_{id}=1)=E(q_{id}|q'_{id}>\tau_d)=E(q'_{id}|q'_{id}>\tau_d)+x_{id}(\beta-\beta')=\tau_d+x_{id}(\beta-\beta')+\lambda(I'_{id}) \quad (8)$$

The above derivation is identical to equation 4, except now the observables x_{id} directly enter the test outcomes equation, even after conditioning on the propensity to test. In other words, the model implies that if observables x_{id} continue to have explanatory power after conditioning on the propensity I_{id}, then physicians are not weighting those observables in the manner that would maximize the incidence of positive tests.

How can we rule out the possibility that untested patients with a given set of observables are known based on unobservables not to have a PE? The function $\lambda(\cdot)$ reveals this information given the assumptions we have made about the distribution of unobservables. This point can be seen most directly by rearranging equation 8 into the form of the first line of equation 4. That is, we can write:

$$E(Z_{id}|Test_{id}=1)=\alpha_d+x_{id}\beta+h(I'_{id}) \quad (9)$$

where $h(I'_{id})=\lambda(I'_{id})-I'_{id}=E(\eta_{id}|\eta_{id}>-I'_{id})$. Written this way, $\alpha_d+x_{id}\beta$ reflects the average return for all patients in the population with this set of observables and $h(I'_{id})$ reflects the fact that the more patients one tests (as one lowers τ_d and thus raises I'_{id}), the more one moves down the marginal benefit curve and tests patients who are likely to have a lower test yield given unobservables.

[13]Note that an analogous argument implies that it is without loss of generality to allow testing thresholds to vary with observables. That is, suppose $\tau_d=\overline{\tau}_d+x_{id}\gamma$. Then we can replace β' with $\beta''=\beta'-\gamma$. In other words, the hypotheses that physicians test patients with a given observable more because they believe those patients are more likely to test positive and that physicians test patients with a given observable more because physicians have a lower testing threshold for patients of that type are empirically indistinguishable.

4.3 Identification

Equation 8 shows that test yields among tested patients depend on physician thresholds (τ_d), allocation of tests to patients ($x_{id}(\beta - \beta')$) and a selection term. As is typical for Heckman selection models, the selection term $\lambda(\cdot)$ can be identified using functional form restrictions, but it would be desirable for $\lambda(\cdot)$ to be semiparametrically identified. We lay out below how semiparametric identification is possible in our setting and how our identifying assumption differs from that used in CS due to the possibility of misallocation.

The CS model is essentially the model we outline in Section 4.1—the one difference is that the dependent variable in equation 4 of our model is whether a patient tested positive rather than an estimate of the causal treatment effect for that patient. In the CS model, identification comes from the fact that x_{id} only enters the test outcome equation (i.e. equation 4) via $\lambda(I_{id})$. In that model, x_{id} are excluded from directly entering the test outcomes equation and we can think of them as instrumental variables which aid in the estimation of $\lambda(\cdot)$, parallel to the standard instrumental variables identification in Heckman selection models (e.g. Mulligan and Rubinstein 2008). This restriction is no longer valid if physicians incorrectly assess the PE risk associated with some observable comorbidities and demographics x_{id}. In the model with misweighting, equation 8 above shows that x_{id} directly enters the test outcomes equation with coefficients that are not known from estimating the equation governing selection into testing.

In order to generalize the model to the case where doctors fail to appropriately weight observable risk factors in deciding whom to test, we consider an additional set of exclusion restrictions.[14] We exploit the fact that τ_d can be directly estimated for physicians testing patients we can identify as marginal.[15] Marginal tested patients are those with the lowest observed values of the testing propensity I'_{id} who are still tested. We estimate the average probability of a positive test among these marginal tested patients. For these patients who are "just barely worth testing," the observed probability of a positive test reveals the threshold at which doctors are willing to test.

Formally, since η_{id} is bounded with full support, there exists some value of the propensity in the testing equation $\underline{I}$ such that patients are only tested for $I'_{id} > \underline{I}$. For those marginal tested patients with $I'_{id} \to \underline{I}$, we know the realization of η_{id} is just barely sufficient to tip these patients across the testing threshold, so that $h(\underline{I}) = E(\eta_{id} | q'_{id} = \tau_d) = -\underline{I}$. Since $\lambda(I_{id}) = I_{id} + h(I_{id})$, it follows that $\lambda(\underline{I}) = 0$ for these marginal tested patients.

Let QQ_d denote the average rate of positive tests Z_{id} among tested marginal patients for doctor d; taking the expectation of equation 8 yields:

$$QQ_d = \tau_d + E_{m,d}\left(x_{id} | Test_{id} = 1\right)\left(\beta - \beta'\right) \quad (10)$$

[14] CS discuss the identification strategy outlined in this paragraph and consider it as a robustness check, but do not directly use it when estimating their model.

[15] More precisely, τ_d is known modulo a misweighting adjustment we spell out below.

In the equation above, $E_{m,d}(x_{id}|Test_{id}=1)$ denotes the expectation of x_{id} only among doctor d's tested marginal patients m. The likelihood of a positive test for those tested patients with the lowest testing propensities is given by the physician's threshold τ_d plus an adjustment for the fact that the actual likelihood of a positive test for these patients differs from physician's beliefs because $\beta \neq \beta'$. This calculation provides an exclusion restriction—after subtracting the average yield among a doctor's marginal tested patients from both sides, doctor fixed effects are excluded for those physicians in equation 8. A more detailed derivation of this result is in Appendix E.

This exclusion restriction also suffices to identify $\lambda(\cdot)$. Intuitively, suppose that by studying marginal tested patients, we uncover multiple physicians with identical thresholds τ_d. These doctors may still differ in their propensity to test for identical observables $\theta_d = \alpha_d - \tau_d$, because they may treat patient populations with different PE risk α_d. After conditioning on τ_d, any remaining doctor-level variation in test outcomes must be explained by differences in patient risk α_d, and the functional form relating α_d to test outcomes will flexibly identify the shape of the $\lambda(\cdot)$ function.

For example, suppose we see many doctors who all share the same test threshold. Some of these doctors test more patients than others because their patient population is riskier (higher α_d). If a small increase in a patient's test probability predicts a large increase in test yield among tested patients (correlating to a much higher α_d), this implies that risk factors observable to the doctor but not observable by the econometrician must heavily influence test yields and thus testing choices. Because a small change in testing behavior correlates with a large change in patient risk, $\lambda(I)$ will be steeply sloped. Technically, this result arises because the density of η_{id} will be smaller as dispersion in η_{id} increases, placing fewer patients in a given neighborhood of the doctor's threshold. Alternatively, if the distribution of unobserved PE risk is less dispersed, i.e. unobservables exert less influence on testing decisions and test outcomes, a given increase in a patient's test probability will predict a smaller increase in test yield among tested patients (correlating to a smaller increase in α_d), implying that $\lambda(I)$ will be relatively flat.

In addition to the validity of the exclusion restrictions, the other crucial identifying restriction underlying this estimation approach is the ignorability assumption: η_{id} is additively separable and i.i.d. across doctors and patients. The ignorability assumption implies that the function $\lambda(\cdot)$ is the same for different doctors and patients. If this assumption were violated and η_{id} were distributed differently across doctors, the function $\lambda(\cdot)$ could be doctor-specific. In Section 6.2, we consider one such model and show that it does not materially impact our results.

In our baseline model, the ignorability assumption contributes to the identification of test yield among currently untested patients. What would happen to test yields if a doctor lowered his threshold (from τ_0 to τ_1) and tested more patients on the margin? The definition of the threshold immediately tells us the new test yield among marginal patients (τ_1). The fact that $\lambda(I_{id})$ embeds information about the entire distribution of η_{id} allows us to infer test outcomes among inframarginal patients as well. Ideally, this variation is "in sample" in the sense that we observe other doctors with a threshold as low as the value we would like to

simulate and can trace out how test yields relate to test probabilities for patients of those doctors. In our baseline model, ignorability implies that this function is the same for all doctors and the threshold varies the y-intercept in (test probability, test yield) space.

The identification of misweighting also relies on the ignorability assumption. The ignorability assumption implies that if doctors were optimally assessing PE risk, any two conditions with the same β' weight in the testing equation should induce the same change in the fraction of positive tests amongst tested patients, holding all other comorbidities and testing thresholds constant. If two conditions with the same β' weight in the testing equation lead to different changes in the fraction of positive tests, then we identify misweighting; we conclude the risk factor that induces the larger increase in positive tests is underweighted relative to the other factor. The slope of the function $\lambda(\cdot)$ with respect to known variation in α_d pins down how x_{id} should impact test outcomes Z_{id} given β'—so we can in principle identify misweighting even with just a single x variable. This strategy echoes the logic of the reduced form evidence on misweighting presented in section 3.5, but the additional structure allows us to make more detailed comparisons of weighting and risk across conditions, after accounting for differences in patient risk and testing thresholds across doctors.

Empirically, the ignorability assumption may be undermined if the distribution of unobserved patient PE risk differs across conditions. For example, if fewer patients with the risk factor that appears to be under-weighted present to the ED with the relevant PE symptoms (e.g. chest pain, shortness of breath, elevated heart rate), then it may be that physicians are already testing every patient in the relevant at-risk population. This assumption is directly analogous to the standard exogeneity assumption used in virtually all structural models; e.g. just as discrete choice models assume that observed product characteristics are independent of the error term, our misweighting model is identified by assuming that specific observed characteristics are not systematically related to unobserved determinants of PE risk.

With unlimited data, we could relax the ignorability assumption. If for every set of x_{id}, we observed sufficient variation in doctor testing choices for patients of that type, we could directly estimate the distribution of η_{id} conditional on x_{id} and check whether testing more patients with a specific underweighted factor leads to more positive tests. In other words, we could allow the function $\lambda(\cdot)$ in our model to be a different function for every set of x_{id}. In practice, we estimate the overall degree of selection on unobservables (assuming η_{id} does not depend on x_{id}), but we lack sufficient data to estimate a separate distribution for each set of x_{id}.

An additional subtlety of our estimation approach is that many doctors test only a small number of patients, so we do not necessarily observe marginal patients for all doctors. Given the ignorability assumption, we can still identify $\lambda(\cdot)$ from the doctors for whom we *do* observe marginal patients, and thus determine τ_d for other doctors.

4.4 Estimation of the parametric model

Let us now specify precisely how we estimate the structural model outlined in the previous sections. Define $\theta'_d = \alpha'_d - \tau_d$. Plugging our specification for the probability of a positive test from equation 5 into the testing equation 2 yields the final form of the testing equation:

$$Test_{id}=1 \leftrightarrow x_{id}\beta' + \theta'_d + \eta_{id} \geq 0 \quad (11)$$

These assumptions yield a binary choice model of testing. In our baseline specification, we assume that η_{id} is i.i.d. across doctors and patients with a parametric distribution we describe below. Thus, patients' ex ante risk distributions may have different means ($x_{id}\beta + \alpha_d$) but are assumed to be otherwise identically distributed. In section 6, we estimate versions of the model which (separately) relax the parametric assumption and allow for heteroskedasticity across doctors in the distribution of patient PE risk.

The most common parametric assumptions in binary choice models—normal and logit—are inconsistent with our model because q_{id} must lie between 0 and 1. Instead, we assume that each η_{id} is drawn from a two parameter distribution which is a mixture of a Bernoulli and a uniform distribution. With probability $1 - p$, $\eta_{id} \sim U[-\eta, \eta]$ and with probability p, $p_{id} \sim U[\upsilon - \eta, \upsilon + \eta]$. Intuitively, this distribution captures the idea that most patients are not candidates for a CT scan. A small fraction of patients p present with symptoms of PE such as chest pain and given those symptoms, there is a range of ex ante risks parameterized by η. We assume that patients are never tested unless they receive the shock υ (i.e. unless they present with PE symptoms).

In addition to these clinical reasons, there are several methodological advantages to this distribution. Among bounded distributions, a uniform distribution is attractive because it leads to a particularly tractable linear selection term $\lambda(\cdot)$. The mixture distribution has two methodological advantages over a pure uniform: firstly, if $p = 1$ (the uniform case), the estimated variance of η is so large that it implies $q_{id} < 0$ for some patients, which is inconsistent since q_{id} is a probability. Secondly, since testing is a low probability event, a uniform distribution would imply that more precise information (a higher variance of η_{id}, meaning that doctors have more private information about test outcomes) leads doctors to test more everything else held equal; the mixture distribution allows for the possibility that more precise information leads to less testing. This second point is especially relevant in the heteroskedastic model considered in Section 6.2 where the variance of η_{id} is allowed to vary across doctors. To demonstrate that our results are not driven by this specific choice of parametric distribution, we also estimate the model semiparametrically as a robustness check in Section 6.3.

In Appendix E, we show that this distributional assumption implies:

$$Pr\left(Test_{id}=1\right) = \max\left\{0, \frac{p}{2} + \frac{p\left(I'_{id}+\upsilon\right)}{2\eta}\right\} \quad (12)$$

where $I'_{id}=x_{id}\beta'+\theta'_d$. Estimation of this equation by non-linear least squares allows us to recover $\hat{\beta}'=\beta'\frac{p}{2\eta}$ and $\hat{\theta}'=\frac{p}{2}+\frac{p(\theta'_d+\upsilon)}{2\eta}$ which we use to construct an estimate of the testing propensity $\tilde{I}'_{id}=\frac{p}{2}+\frac{p(I'_{id}+\upsilon)}{2\eta}$.

Following the steps outlined in the previous section, the testing threshold parameters τ_d can be recovered from a regression of test outcomes (i.e. positive or negative for detecting PE) on doctor fixed effects, controlling for the propensity I'_{id} estimated from the testing equation. Note that under the parametric assumptions we have made so far,

$E(\eta_{id}|\eta_{id}>-I'_{id})=\frac{\eta-I'_{id}+\upsilon}{2}$. As shown in more detail in Appendix E, this implies that:

$$E\left(Z_{id}|Test_{id}=1\right)=\tau_d+x_{id}\left(\beta-\beta'\right)+\frac{\eta\tilde{I}'_{id}}{p} \quad (13)$$

As discussed in section 4.3, we avoid relying solely on functional form to identify the coefficient on $\tilde{I}'_{id}$ by estimating τ_d directly for doctors with tested marginal patients based on the observed average rate of positive tests among those marginal patients, $\widehat{QQ}_d$. We define marginal patients as patients in the first decile of $\tilde{I}'_{id}$ among tested patients; this definition is conservative from the standpoint of detecting overtesting since more restrictive definitions (e.g. the first percentile) will tend to lead to lower estimated thresholds. We show in Appendix Table A.1 how our estimates change for alternative definitions of marginal patients. As expected, we estimate lower τ_d (and thus more implied overtesting) using more restrictive definitions.

Subtracting $\widehat{QQ}_d$ from both sides of equation 13 yields:

$$Y_{id}=\left(1-M_d\right)\tau_d+\frac{\eta\tilde{I}'_{id}}{p}+X_{id}\left(\beta-\beta'\right)+\varepsilon_{id} \quad (14)$$

where $Y_{id}=Z_{id}$ for doctors with no tested marginal patients and $Y_{id}=Z_{id}-\widehat{QQ}_d$ for doctors with marginal patients, M_d is an indicator for whether a doctor has marginal patients, $X_{id}=(x_{id}-E_{m,d}(x_{id}))$ for doctors with marginal patients and x_{id} for doctors with no marginal patients.

One could estimate equation 14 in two steps—first, estimating the model among doctors with marginal patients with doctor fixed effects omitted to recover $\frac{\eta}{p}$ and $\beta-\beta'$, and then estimating the model among doctors with non-marginal patients to recover the full set of doctor thresholds τ_d. Because fixing either $\frac{\eta}{p}$ or $\beta-\beta'$ would be sufficient to identify

equation 14 for doctors with non-marginal patients, estimating the model jointly for all doctors uses additional information about the relative value of the parameters for doctors with non-marginal patients; this increases the precision of the estimates but has little impact on the magnitude of the coefficients.

Least squares estimation of equation 14 will allow us to recover the constant $\frac{\eta}{p}$ and doctor fixed effects τ_d for non-marginal patients which, when combined with our estimates for marginal patients from $\widehat{QQ}_d$, can be used to recover the full distribution of estimated $\hat{\tau}_d$.

The distribution of $\hat{\tau}_d$ combines both the true underlying variation in τ_d and estimation error from the fact that each τ_d is imprecisely estimated. To correct for estimation error, we apply an "empirical Bayes" technique to recover moments of the true underlying distribution of τ_d. Our approach is described in detail in Appendix F.[16] Unlike more standard estimators (such as Kane and Staiger 2008), this technique is robust to the fact that we observe only a small number of observations per doctor and makes no distributional assumptions about either the true distribution of τ_d or the estimation error. The true distribution cannot be nonparametrically identified, but we can recover moments of that distribution; we report the mean and standard deviation. Simulation results do require us to recover a posterior estimate of τ_d for each doctor, and for these exercises we impose a further assumption that τ_d is log-normally distributed as described in Appendix F.

5 Results

In this section, we report results of the estimation strategy described in section 4.4 above. First, we describe the recovered distribution of physician testing thresholds and test how physicians' training and practice environment are related to testing intensity. Then, we report results on which risk factors are under- and over-weighted in physicians' risk assessments relative to the weighting that would maximize detection of positive tests and consider possible clinical explanations for these patterns. Finally, we simulate how variation in test thresholds and the presence of misweighting affects physicians' test yields.

5.1 Distribution and correlates of physician testing thresholds

After estimating the model laid out in Section 4 and applying the empirical Bayes adjustment, we find the mean value of τ_d is 0.056 and and the standard deviation is 0.054.[17] In other words, the average doctor is willing to test a patient provided the doctor's estimate of the probability of a positive test exceeds 5.6%. Note that this positive test rate includes tests which detect actual PEs and false positives. The standard deviation of 0.054 suggests that there is a large amount of heterogeneity across doctors in their testing thresholds, with some doctors testing almost all patients displaying the relevant symptoms, and other doctors testing only patients with very substantial PE risk. Considering that the overall test yield in

[16] We use quotation marks since our procedure is not a traditional empirical Bayes appraoch: we do not derive our estimator as the posterior of any specific distribution.

[17] Note that of course this would not be consistent with a normal distribution since in this case $\tau_d > 0$ for all doctors or they would test every patient. In our welfare exercises we assume a log-normal distribution.

our sample is only 6.9%, it is likely that this variation in testing thresholds may affect testing decisions for many patients.

We next consider regressions of the estimated testing thresholds $\hat{\tau}_d$ on doctor, hospital and regional characteristics to explore the determinants of practice style. Specifically, we regress $\hat{\tau}_d$ on variables capturing doctor experience (the number of years since the doctor graduated from medical school), whether the medical school the doctor attended is ranked in the top 50 for research or primary care by US News & World Report, whether the hospital where the physician practices is a for profit hospital or an academic hospital, regional medical spending, the state tort environment, and average income in the region.

We consider OLS estimates as well as FGLS estimates which take into account the estimation error in the dependent variable τ_d.[18] For each specification, we consider models with and without hospital fixed effects. Including hospital fixed effects to identify the impact of within-hospital variation in physician characteristics obviates the concern that our model omits unobserved differences in the cost of testing at the hospital level. For example, there may be variation in the opportunity cost of testing, depending on whether the CT scan is used to capacity. This heterogeneity will be absorbed into the hospital fixed effect.

Table 3 reports the results. We find that doctors in higher spending regions have lower testing thresholds, i.e. they are more likely to test low risk patients. A 10% increase in regional spending, as reported by the Dartmouth Atlas, is associated with a 0.4 percentage point decline in testing thresholds, significant at the 1% level. This finding provides empirical support for the hypothesis that high spending regions are providing lower marginal value, “flat of the curve” medical care.

We also find evidence that more experienced doctors have higher testing thresholds: a 10-year increase in doctor experience is associated with 0.7 percentage point higher testing thresholds, significant at the 1% level. This relationship persists after controlling for hospital fixed effects, suggesting that even within the same institution, more experienced doctors are less likely to test low-risk patients. Unfortunately, we do not observe enough testing decisions per physician to estimate the model with year-specific testing thresholds for each physician, and as a result we cannot disentangle cohort and experience effects. Our finding stands in contrast to the result in Cutler et al. (2013) that older physicians are more likely to recommend agressive treatment for cardiac patients. One explanation for this difference may be that older physicians were trained before the broad diffusion of modern CT scans which are used to diagnose PE, and so may be more likely to rule out pulmonary embolism on the basis of clinical presentaation.

Many factors predicted to influence care quality, such as the quality of the physician’s training, the financial structure of the hospital (for profit or otherwise), its status as an academic institution, and the income of the patients served, have no significant relationship to testing thresholds. Estimates relating physician’s medical school rank to testing thresholds

[18]The FGLS estimates are based on Lewis and Linzer (2005), where the error term consists of both a homoskedastic e_{id} with unknown variance and a heteroskedastic component with known variance. The heteroskedastic component arises from the estimation error in $\hat{\tau}_d$ which is in turn recovered from estimation of equation 14.

are imprecisely estimated, with the upper bound of the 95% confidence interval at a 1.2 percentage point higher threshold for those attending a top 50 research institution. Point estimates suggest slightly higher thresholds for academic hospitals and lower thresholds among for-profit hospitals, but the 95% confidence intervals bound the differences in average thresholds to less than one percentage point.

Finally, exploiting cross-sectional variation in enactment of tort reform, including joint and several liability and malpractice damage caps, we find no consistent relationship between the malpractice environment and testing thresholds. The FGLS estimates point to a significant, negative relationship between testing thresholds and malpractice damage caps, which would be the opposite prediction of theory suggesting physicians are more likely to test low-risk patients in states with damage caps. The coefficient is much smaller in magnitude and no longer statistically significant in the OLS specification. Our lack of power to estimate year-specific testing thresholds precludes us from undertaking a difference-in-differences analysis of malpractice law.

Given the large estimated variation in τ_d, with a standard deviation of 0.054 after adjusting for statistical noise, observed factors can explain only a small fraction of the estimated variation in physician practice style. This observation implies that policy responses targeted at reducing testing rates in specific hospital types (e.g. for profit hospitals) or policies aimed at raising the qualifications of emergency department doctors are unlikely to lead to substantial reductions in testing variation. Instead, focusing on policies which target the decision-making process rather than physician credentials or practice environment may have greater scope for reducing heterogeneity in practice style. This parallels the finding in the teacher fixed effects literature that there is substantial variation in teacher productivity not explained by teacher credentials or other observable factors (Jackson et al. 2014).

5.2 Identifying misweighted comorbidities

Next, we explore physicians' misweighting of observable PE risk factors. As outlined in section 4.2, we focus on measuring aggregate misweighting: factors which appear to be systemically under- or over-weighted in physicians' assessments of patient PE risk. The model implies that physicians are overweighting a given risk factor if they are substantially more likely to test a patient with that factor (holding constant other observable patient characteristics), but this variable does not yield a commensurate increase in the rate of positive tests among tested patients. The evidence of both under- and over-weighting suggests that physicians could perform the same total number of tests but detect more PE cases, if they improved targeting of the tests by applying different weights to important risk factors.

Results are reported in Table 4 and Appendix Table A.2. For each risk factor in our model, column 1 reports the marginal effect of this variable on testing probability based on the coefficient β' from the testing equation (cf. equation 5). Column 2 reports the estimated error in physicians' assessment of the PE risk associated with each comorbidity, implied by how the weights attached to each comorbidity in their testing decisions compare to the conditional influence of each comorbidity on test outcomes (cf. equation 13). Finally,

columns 3 and 4 report the standard error and t-statistic on estimated misweighting, respectively. Variables are sorted by their t-statistic in this table.

Given our nonlinear model, the reported marginal effects in column 1 hold for all patients for whom $\tilde{I}'_{id}>0$, which is true for the average patient in our data. (Marginal effects are zero for patients with negative values of $\tilde{I}'_{id}$.) All included risk factors are binary variables; variables with the most misweighting will have the largest absolute value of misweighting reported in column 2. We report robust standard errors that don't account for estimation error in the testing propensity index $\tilde{I}'_{id}$, although this adjustment would be very small given the large sample of patients identifying $\hat{I}'_{id}$.

We find evidence of substantial under- and over-weighting of key risk factors, relative to the weights that would maximize test yields. Comparing physician's implied prediction of PE risk for each patient with the estimated actual risk, we find that physicians appear to be misestimating a patient's probability of a positive test by 2.3 percentage points on average, accounting for all comorbidities and averaging the absolute value of each patient's aggregate misweighting to include both under- and over-estimates. This degree of misestimation has the potential to affect testing decisions for many patients.

Investigating the specific conditions that drive the aggregate misweighting, we find that doctors appear to react strongly to patients' clinical symptoms, overtesting patients with clinical conditions that may mimic the symptoms of PE, while discounting the importance of known PE risk factors from the patient's medical history. We cannot distinguish in this setting whether the apparent overattention to symptoms rather than comorbidities is driven by inadequate information in the emergency care context about patient's medical history or by mistaken beliefs about the PE risk associated with each factor. Future research could study whether high quality electronic medical records mitigate this problem by providing timely information about relevant medical history or whether tailored decision support might help guide physicians' assessment of patient PE risk.

The strongest evidence of underweighting comes from physicians' implicit estimate of the PE risk associated with a recent inpatient admission history. While immobilization is a commonly known risk factor for PE, popular risk scores highlight the role of recent surgery but do not broadly include other types of hospitalization. Perhaps as a result, we see evidence that physicians have adequately increased testing rates for patients with a recent surgical history, but do not place sufficient weight on recent hospital admissions that did not include a surgical procedure. The marginal effect reports that physicians are 0.9 percentage points *less* likely to test a patient with a prior inpatient admission within the past 30 days, implying that doctors have underestimated these patients' PE risk by 11 percentage points after accounting for the role of other observed comorbidities.

In addition, several specific cancer diagnoses and a history of PE or the related condition deep vein thrombosis show evidence of substantial underweighting, suggesting that physicians are failing to adequately consider these risks when assessing a patient for PE.[19] For all but one of these conditions (metastatic cancer), physicians are indeed more likely to

test patients with the observed condition, holding constant other patient risk factors, but the response is not adequate given the large influence of this preexisting condition on the current risk for PE. This pattern is occurring despite the fact that both cancer treatment and history of PE or deep vein thrombosis are two of the seven risk factors in a popular PE risk-scoring algorithm known as the Wells score. This suggests that physicians are continuing to under-respond to these critical risk factors despite their recognized role in PE risk.[20]

A few other risk factors also show evidence of significant underweighting, including rheumatoid arthritis, obesity and paralysis, all of which are known risk factors for PE documented in the medical literature, although not explicitly included in popular risk scoring algorithms. A complete list of underweighted risk factors is reported in the top panel of Table 4.

A number of different conditions that mimic the symptoms of PE appear on the list of over-weighted comrbidities: these are conditions where test yields are predicted to improve if physicians became less likely to test patients with these particular conditions. The three conditions with the most significant evidence of overweighting (i.e. atrial fibrillation, chronic obstructive pulmonary disease, and ischemic heart disease), have chest pain and difficulty breathing as hallmark symptoms; these are also key clinical symptoms of PE. Severe depression often manifests in the emergency department context with somatic symptoms of chest pain and shortness of breath as well. Patients who visit the emergency department with an exacerbation of another previously diagnosed condition could be suspected of having PE due to similar symptoms and thus may be tested at a higher rate even though our data suggests they are not at higher risk of PE, holding constant their other risk factors. Given that these other conditions must have been diagnosed prior to the emergency department visit in order to be included on our comorbidity list, physicians should be aware of them at the time they are evaluating the patient for PE. Of course, failure to take an appropriate medical history or limited access to patients' prior health records could hinder evaluation and contribute to the observed overweighting of these conditions.

Turning to demographic variables, we find evidence that black patients are under-tested. They are less likely to be tested for PE than non-black patients, despite the fact that they are at higher risk of PE. Given the structure of our model, these differences in testing patterns of black and white patients cannot be explained by differential sorting to physicians, since we have controlled for differences in physicians' testing thresholds. This finding provides new empirical support for the concern about racial disparities and possible provider prejudice in medical treatment (cf. Nelson 2002). The result stands in contrast to results from Chandra and Staiger (2010) that applied a related analytic framework to a different clinical setting and found that while blacks receive less treatment for heart attacks, differences were fully explained by their lower benefits from treatment. In the setting of testing for PE, differences

[19]Prostate cancer, metastatic cancer, endometrial cancer and colorectal cancer all have significant underweighting.
[20]Whether the underweighting of these risk factors is driven by failure to adhere to Wells' score criteria or whether the Wells score inadequately weights these risks is not something we can directly assess in our data. Complete calculation of the Wells' score would require information that is difficult to observe in claims data or even retrospective study of patient charts. For example, the most highly weighted factor in the score is the physician's clinical opinion that PE is the most likely diagnosis, or equally likely to the other possible diagnosis.

in test yields do not explain disparities in testing rates. Notably, these disparities are arising among patients who have all arrived at the emergency department for evaluation by a physician with access to a CT scanner, and all carry Medicare insurance coverage, although they may differ in their subscription to wrap-around private insurance.

Taken together, these results suggest that misassessments of the clinical risk associated with preexisting comorbidities may lead to substantially diminished test yields. It is possible that physicians could detect more PE cases while performing a similar number of tests, by adjusting the targeting.

An alternative explanation for these patterns of apparent misweighting would be that the value of detecting PE differs for patients with these varying risk factors. For example, if the value of detecting PE were substantially lower in patients with a recent hospital admission or a cancer diagnosis, that could explain the apparent underweighting. Conversely, if the value of detecting PE were higher for patients with ischemic heart disease, COPD or atrial fibrillation, then that could also help rationalize the observed testing behavior. We find no obvious link between most of these conditions and the value of PE detection. In fact, our results on age-related risk suggests that physicians are undertesting younger patients, for whom the value of PE detection should be particularly high, since they have a longer life expectancy and accordingly higher value of statistical life. One exception in which a lower value of treatment may explain the observed results is Alzheimer's disease; this appears in our list of underweighted conditions, but may reflect the lower value of treating pulmonary embolism among patients with this severe, progressive disease.

5.3 The impact of threshold variation and misweighting on test yields

To quantify the role that testing thresholds and misweighting play in the observed patterns of testing behavior and test yields, we return to the graph of physician testing rates and test yields. Now, rather than binning physicians by the average fraction of patients tested as we did in Figure 2, we bin physicians by the structural analogue: the average estimated testing propensity $\widehat{\tilde{I}}'_{id}$ across their patients. Recall the observation from the reduced form analysis in section 3.4 that physicians with the highest average testing rates also had the lowest test yields. This downward sloping relationship is what we would expect to find if heterogeneity in τ_d were the primary driver of observed variation in testing rates across doctors.

We can explore this hypothesis more formally by using our model to simulate what the relationship between average physician testing propensities and positive test rates would have been if all doctors had the same testing threshold. We simulate testing decisions and test outcomes under a counterfactual where τ_d is held constant across doctors, at the estimated average value $E(\tau_d) = 0.056$. Details of this simulation are provided in Appendix G.

Results of this exercise are pictured in Figure 3. The open circles depict the downward sloping relationship between physicians' average testing propensities and their test yields in our observed data. As we suggested earlier, if all doctors had the same testing threshold, the remaining variation in doctors' average testing propensities would be driven by differences

in patient risk of PE. As a result, the relationship between doctors' average testing propensities and their test yields would become upward sloping over most of the domain. The solid square markers display the results of this simulation in Figure 3. Now the doctors with higher testing rates are those with the highest risk patients; these doctors test the greatest fraction of their patients and experience the highest test yields, as evidenced by the upward slope in the simulated plot.[21]

Finally, we investigate how misweighting impacts this relationship between testing propensity and test yields. We simulate the counterfactual relationship between physicians' average testing propensities and test yields that would be observed if there were no heterogeneity in testing thresholds *and* no misweighting of observable risk factors. Eliminating misweighting should increase the test yield for all values of the testing propensity index by improving the targeting of PE CT tests. Details of the simulation exercise are described in Appendix G.

Results of this simulation are pictured in Figure 3 and plotted with the X-shaped markers. We see that for every decile of physicians' average testing propensity, the predicted test yield is higher in the simulation with no misweighting than was observed in both our actual data or the simulation that only eliminated threshold variation. We predict more detected positive tests if physicians attached appropriate weights to observable risk factors, and the increase is largest at lower testing propensities. (We quantify the precise increase in test yields and their welfare consequences in section 7.3.) Inframarginal patients are likely to be tested even with misweighting, but the set of marginal patients changes—some patients who are less likely to test positive are no longer tested and others who were previously not tested but have a higher likelihood of testing positive are now tested. This exercise suggests that misweighting is a substantial contributor to low test yields, and attention to better targeting of testing resources is warranted, rather than focusing solely on reducing variation in testing rates.

6 Robustness

The results discussed in the previous sections depend on a number of modeling assumptions. Two crucial assumptions underly our identification arguments: first, that we can identify marginal tested patients and use their test yields to reveal physician's test thresholds; second, that the restrictions we assume for the η_{id} term, the factors influencing testing choices that are observable to the doctor but unobservable to the econometrician, are valid. In our baseline specification, we assume that η_{id} is i.i.d. across patients and doctors and follows a specific parametric distribution. In the robustness checks described below, we test the sensitivity of our results to these assumptions. Specifically, we consider the robustness of our results to varying the set of included covariates and the definition of the marginal tested patients; we estimate a version of our model where the variance of η_{id} is allowed to vary

[21] If we graphed testing propensities vs. simulated rates of positive tests at the individual patient level, fixing $\tau_d = E(\tau_d)$, our model implies that the resulting relationship would be monotonic. Because we are aggregating to the physician level in the figure, this relationship also depends on the variance in testing propensities for a given physician; the slight non-monotonicity at the lowest deciles arises because doctors with the lowest average testing propensities have more heterogeneous patients (driven by variation in observed comorbidities x_{id}) than those in adjacent deciles. At these low average testing propensities, higher variance in I_{id} is associated with more positive tests amongst tested patients due to the convexity of the relationship between I_{id} and positive testing rates at the individual level.

flexibly across doctors; and we estimate a semiparametric model where η_{id} is once again assumed to be homoskedastic but now with an arbitrary distribution.

6.1 Stability of results to inclusion of alternate patient controls

In the spirit of Altonji et al. (2008), we explore the sensitivity of our results to the set of included variables to assess potential bias from unobservable risk factors. The rationale for this exercise is that omitting the variables x_{id}^{omit} from the baseline specification could generate heteroskedasticity, if the resulting error term $\eta_{id}^{'}=\eta_{id}+x_{id}^{omit}\beta$ is not independently and identically distributed across doctors and patients. If this heteroskedasticity substantially changes our estimates of the distribution of τ_d or the degree of misweighting for the remaining variables, this might suggest that including additional unobserved variables would change our estimates further.

Recall that we rely on comorbidities to identify the patients the doctor is just indifferent between testing and not testing, and then calculate test outcomes among that group to identify thresholds for physicians with marginal patients. In addition to testing robustness to heteroskedasticity in the error term, varying the set of included variables will also change the set of patients identified as marginal (i.e. just barely worth testing given their physician's threshold). As we remove comorbidities from the analysis, we are less able to isolate the marginal patients and may include more inframarginal patients in the group used to identify doctor's testing thresholds. To show exactly how varying the definition of marginal patients impacts the analysis separately from heteroskedasticity, we also consider explicitly varying our threshold quantile for which patients count as marginal.

The baseline model reported above included four main classes of patient level risk factors: PE specific risk factors, chronic condition warehouse comorbidities, Elixhauser comorbidities, and patient demographic variables. Because some variation in comorbidities is required to appropriately identify this model, we retain the PE specific risk factors and the chronic condition warehouse comoribidities throughout, and test the stability of our findings to excluding the Elixhauser comorbidity set and the vector of demographic variables. Results from this exercise are reported in Table 5; the empirical Bayes correction has been applied before reporting the mean and standard deviation of physician's testing thresholds.

The mean estimated value of physician's testing thresholds ranges between 5.6% and 6.6%, and shows evidence of substantial dispersion in all models. The standard deviation of τ_d ranges between 3.9% and 5.4%, depending on the set of included patient risk factors. Dropping covariates does appear to increase the value of the estimated mean τ_d although the range of values across specifications is only one quarter of the estimated across-doctor standard deviation. If including additional covariates would cause estimates of τ_d to decrease, this suggests that our results may be conservative with respect to the amount of overtesting. Controlling for the full set of risk factors also appears to increase the variance in estimated testing thresholds, providing suggestive evidence that the observed variation in thresholds is not driven by the exclusion of unobserved risk factors from the model. In all of these cases, variation in testing thresholds is sufficient to imply large differences in testing probabilities for identical patients depending on which doctor they visit.

It is not surprising that the mean τ_d increases when we exclude covariates. When we exclude comorbidities from the sample, we make it more difficult to identify accurately the marginal tested patients, and may end up including more non-marginal patients in this calculation. These non-marginal tested patients will have higher average test yields, and so will push up our estimated test thresholds. To examine more directly the sensitivity of our results to the definition of marginal patients, we explicitly vary this definition in Appendix Table A.1. We include all the baseline covariates but vary the quantile of the testing propensity cutoff below which patients are defined as marginal. Less stringent definitions of marginal patients than in our baseline results recover a larger average value of the physician threshold as predicted and more stringent definitions recover a lower value, suggesting our results are conservative with respect to the degree of overtesting to the extent that with more data (or more covariates) we could better identify those patients who were truly marginal.

All specifications also predict substantial misweighting of included risk factors. The average absolute value of misweighting in physicians' assessment of PE risk ranges from 0.020 to 0.023 percentage points. Perhaps unsurprisingly, the full model which includes all available risk factors as candidate sources of misweighting recovers the largest predicted amount of misweighting. In all cases, misweighting is sufficiently large that it has the potential to change testing decisions for many marginal patients. Appendix Table A.1 reports that varying the definition of marginal patients also does not change the estimated misperception of PE risk.

In results reported in Appendix Table A.3, we find that the specific misweighted factors identified in Table 4 and discussed in section 5.2 continue to show evidence of misweighting of similar direction and magnitude, even as we vary the set of other included comorbidities. For example, the PE risk associated with recent hospital admissions and history of PE or deep vein thrombosis appears significantly underweighted in all specifications; black patients also show evidence of being under-tested in both specifications that include demographic variables. Similarly, a consistent set of conditions shows evidence of overweighting across specifications, including ischemic heart disease, chronic obstructive pulmonary disease and atrial fibrillation. These findings are not sensitive to the choice of other included covariates.

6.2 Estimation with physician-specific heteroskedasticity

Even if our results are not sensitive to dropping some covariates, we might worry that PE risk factors we cannot observe from insurance claims vary systematically across doctors. Differences across doctors in the variance of η_{id} could arise for at least three reasons. First, doctors may differ in their skill at assessing risk factors unobservable to the econometrician. A doctor with more diagnostic skill may have a higher variance in η_{id} across his patients, since he is more discerning in his judgement of which patients should be tested on the basis of clinical presentation and symptoms. Second, doctors may differ in the variance of latent PE risk present in their patient population. A doctor with a more heterogeneous patient population may have a higher variance in η_{id} across his patients. Finally, doctors may simply make "errors" that lead them to deviate from typical practice patterns; a doctor who frequently deviates from his peers' practice patterns in assessing PE risk may have have a

higher variance in η_{id}. The model we develop in this section allows us to isolate differences in physician testing thresholds that are unrelated to possible differences in the variance of η_{id} across physicians.

Recall the assumption we made in Section 4.4 that η_{id} followed a mixture of a Bernoulli and uniform distribution. We maintain the basic shape of the distribution but now allow both the Bernoulli probability and the variance of the uniform distribution to vary across doctors, so that $\eta_{id} \sim U(-\eta_d, \eta_d)$ with probability $1 - p_d$ and $\eta_{id} \sim U[\upsilon - \eta_d, \upsilon + \eta_d]$ with probability p_d.

Following the derivation in Appendix E, the more flexible distributional assumption implies the testing equation takes this form:

$$Pr\left(Test_{id}=1\right) = \max\left\{0, \frac{p_d}{2} + \frac{p_d\left(I'_{id}+\upsilon\right)}{2\eta_d}\right\} \quad (15)$$

From the testing equation above, we can see that heteroskedasticity in η_{id} is identified by the fact that observables are less predictive of testing behavior for doctors with a high variance in η_{id}, i.e. a smaller value of $\frac{p_d}{\eta_d}$. As described in the appendix, the testing equation can be used to estimate $C\frac{p_d}{2\eta_d}$, where C is an unknown scaling constant. For computational tractability given the demands of this more flexible estimation strategy, we randomly exclude half of the physicians from our sample to reduce sample size, and drop the Elixhauser comorbidities and demographic risk factors from our list of included covariates.

With the introduction of heteroskedasticity, the conditional probability of a positive test is given by:

$$E\left(q_{id} | Test_{id}=1\right) = \tau_d + \frac{C}{2}\frac{\tilde{I}'_{id}}{\hat{\eta}_d} + \left(x_{id} - E_d\left(x_{id}\right)\right)\left(\beta - \beta'\right) \quad (16)$$

where $\hat{\eta}_d = C\frac{p_d}{2\eta_d}$ are the variances estimated in the testing equation. Further details of the estimation strategy are provided in Appendix E.

Table 5 reports the results of this analysis in panel 4, which can be compared to results from the baseline model with the same excluded comorbidity set, as reported in panel 3. The mean value of physicians' test thresholds τ_d is slightly higher at 7.0% in the model allowing for heteroskedasticity compared to 6.6% in the baseline model with the same covariates. Estimates of the standard deviation of τ_d are are also higher at 5.1 percentage points in the heteroskedastic model compared to 3.9 percentage points in the homoskedastic model. Thus, the cross-physician variation in testing behavior is not explained by differences in the variance of η_{id} across doctors. This provides reassuring evidence that the assumption of

homoskedasticity in the baseline model was not leading us to overstate differences across physicians in testing thresholds. Finally, the degree of misweighting remains very similar to the original estimates, with the average absolute value of misweighting estimated at 0.021 in the heteroskedastic model compared to 0.020 in the baseline model.

The role of physician diagnostic judgment in driving testing behavior and outcomes was previously explored by Doyle, Ewer, and Wagner (2010). In a natural experiment, they find that physicians from more prestigious residency programs achieve similar patient outcomes at 10–25% lower cost compared to their less skilled peers. One potential explanation for this phenomenon is that physicians from less prestigious schools prefer to administer more low-value care and could achieve the same outcomes at lower cost if they cut back some services. In the language of our model, these less skilled physicians might have lower testing thresholds, i.e. smaller τ_d. A second explanation is that these less skilled physicians just need to use more medical resources to achieve the same quality of care, because they are less accurate in their assessments of ex ante patient risk. In the language of our model, this decreased diagnostic accuracy would correspond to a lower variance of η_{id}, since these less skilled physicians would be failing to incorporate clinical information about patient risk to improve test targeting. Our results suggest that the heterogeneity in measured τ_d across physicians persists even after allowing for heterogeneous variance of η_{id} across doctors. This finding raises the possibility that cost variance across physicians is driven in part by lower marginal value services provided by doctors with lower expected benefit thresholds.

6.3 Estimation of a semiparametric selection model

Next we test whether our results are sensitive to the shape of the distribution assumed for the unobserved component of patient PE risk, η_{id}. We previously imposed a strict distributional assumption, requiring η_{id} to be distributed according to a mixture of Bernoulli and Uniform distributions. Now, we relax this assumption by estimating Equation 11 as a semiparametric binary choice model, using the Klein and Spady (1993) binary choice estimator. This robustness exercise will ensure that differences in testing thresholds observed in the previous sections are not driven solely by the strong distributional assumptions which restricted the functional form of the testing equation and the shape of the selection correction function $\lambda(\cdot)$. To implement the semiparametric model, we return to our original, strong version of the ignorability assumption that η_{id} is i.i.d. across physicians and patients.

Estimation of the semiparametric model proceeds as follows. Let g denote the probability that patient i is tested given index $I'_{id}=x_{id}\beta'+\theta'_d$. The log likelihood is given by:

$$L(\beta, g)=\sum_i \left[Test_{id} \ln g\left(x_{id}\beta'+\theta'_d\right)+(1-Test_{id})\left(1-\ln g\left(x_{id}\beta'+\theta'_d\right)\right)\right] \quad (17)$$

The idea of the Klein-Spady estimator is to approximate g using a “leave-one-out” estimator which predicts the probability of testing for a particular patient, giving more weight to patients with nearby indices I'_{id}. Specifically, we substitute for g using the following function:

$$\hat{g}_{-i,d}=\frac{\sum_{j\neq i}k\left(\frac{I'_{jd}-I'_{id}}{h}\right)Test_j}{\sum_{j\neq i}k\left(\frac{I'_{jd}-I'_{id}}{h}\right)} \quad (18)$$

We use a 4th-order Gaussian Kernel, $k(\cdot)$, and empirically select for the smallest bandwidth h such that g is a monotonic function of the index I'_{id}.

Given the propensity to test index I'_{id} from estimating equation 11 by the Klein-Spady procedure, the next step is to estimate the testing outcome equation. Echoing the derivation in Section 4.2, the probability of a positive test among tested patients is given by:

$$E\left(Z_{id}|Test_{id}=1\right)=\tau_d+x_{id}\left(\beta-\beta'\right)+\lambda\left(I'_{id}\right) \quad (19)$$

where $\lambda\left(I'_{id}\right)=I'_{id}+h\left(I'_{id}\right)$. Because we no longer assume a particular distribution of η_{id}, we now fit the function $\lambda(\cdot)$ flexibly, reporting results with $\lambda(\cdot)$ as a linear function and as a cubic polynomial, and estimate the net benefit equation by OLS.

Note that the Klein-Spady estimator only recovers I'_{id} up to a location and scale normalization. The scale normalization is embedded in the function $\lambda(\cdot)$. We impose the appropriate location normalization so that at the smallest value of I'_{id} among tested patients, $\underline{I}$, we have $\lambda\left(\underline{I}\right)=0$ as shown in Section 4.3.[22]

Estimation of the semiparametric model is quite computationally intensive, and as a result, we maintain the restricted sample size and covariate set also used in the estimation of the heteroskedastic model in the previous section. Each time we construct the likelihood function, we need to construct a jackknife estimate for each observation which is a weighted average across all other observations given our kernel and bandwidth. This is nested within an optimization problem in which we estimate the parameters of our model for a given bandwidth. We then iterate the entire procedure, searching over for the smallest bandwidth that gives a monotonic result.

Results of the semiparametric estimation are reported in Table 5, panels 5 and 6. This semiparametric estimation approach estimates the mean value of τ_d at 6.7% (linear) or 6.6% (cubic), similar to the parametric model estimate of 6.6% in the sample with identical comorbidities. We continue to find a large amount of cross-doctor dispersion in estimated testing thresholds. The standard deviation of τ_d is 5.4% across doctors, compared to 3.9% in

[22]This normalization can be implemented by omitting the constant term from the polynomial $\lambda(\cdot)$ and subtracting a constant $\underline{I}$ from $\hat{I}'_{id}$; thus the resulting polynomial $\lambda\left(I'_{id}-\underline{I}\right)$ will equal 0 for $I'_{id}=\underline{I}$. To avoid sensitivity to outliers, we normalize I'_{id} so that $\lambda\left(\underline{I}\right)=0$ for I'_{id} in the 10th percentile amongst tested patients, which agrees with our definition of marginal patients in Section 4.3.

the parametric model with the same covariates (but interestingly nearly identical to the parametric model with the full set of covariates included). Our assessment of misweighting continues to be highly consistent across models, with an average absolute value of the error due to misweighting at 2.1% in the semiparametric model, compared to 2.0% in the parametric model.

Taken together, these robustness checks, including varying the set of included covariates, allowing for physician-specific heteroskedasticity, and estimating a semiparametric selection model, all suggest that our findings on the dispersion in testing thresholds and amount of misweighting are very stable across alternative modeling assumptions. We find substantial variance in testing thresholds of similar magnitude in all specifications, suggesting that much of the observed variation in testing behavior may be driven by differences in practice styles. Further, doctors are misassessing patient PE risk by similar amounts in percentage point terms across all models.

7 Welfare cost of overtesting and misweighting

We now turn to the welfare implications of the models estimated in the previous sections. In order to assess the welfare cost of overtesting and misweighting, we will need to make additional assumptions about the costs of testing and the dollar-equivalent benefits of detecting and treating a PE. Given these assumptions, we can evaluate whether the observed variation in testing thresholds reflects overuse and compare the welfare cost of overuse to the welfare cost of misweighting. Applying the structure and estimates of our baseline estimation procedure, we perform simulations to determine how welfare would change if doctors behaved optimally from a social standpoint. We begin by simulating worlds with no overtesting but maintaining the observed patterns of misweighting; next, we simulate a world with no misweighting but maintain the observed distribution of testing thresholds. In each case, we decompose the sources of estimated welfare gains into financial costs, medical costs and medical benefits.

This section proceeds first by describing the calibration of the optimal testing threshold τ^*, then exploring the welfare implications of the measured variation in physician testing thresholds, and finally estimating the welfare costs of misweighting the PE risk associated with patient comorbidities. All of the calibrations in this section are implemented in our baseline model as outlined and reported in Sections 4 and 5.

7.1 Calibration of parameters

In order to proceed with welfare calculations, we make several additional assumptions about the costs of testing and the benefits of a positive test. We assess these costs and benefits from a social standpoint; e.g. if some physicians test more due to reimbursement incentives, this would appear in our model as measured heterogeneity in τ_d that deviates from the social optimum we compute below.

If physicians are behaving optimally, they should test a patient if and only if: $NUq_{id} - c > 0$ where NU represents the net utility of detecting a positive test, c represents the cost of the

test and as above, q_{id} denotes the likelihood of a positive test. This yields a socially optimal testing threshold $\tau * = \frac{c}{NU}$ such that physicians should test only if $q_{id} > \tau^*$.

If there were no false positive or false negative tests, the net utility would correspond to the net medical benefits of treating PE minus any financial costs of treatment. However, CT scans, like many other medical tests, can generate both false positive and false negative results (Stein et al. 2006). It turns out that an important cost of overtesting is a consequence of type I and type II errors: overtesting leads to unneeded treatment which can have adverse consequences. Patients with false positive test results receive medical treatment as if they truly had a PE; this treatment will incur medical risks and financial costs without conferring any medical benefit on the patient, since they do not truly have the condition being treated.

Let *fp* denote the likelihood of a false positive, *s* the sensitivity of the test (one minus the probability of a false negative), *MB* the medical benefits of treating a PE, *MC* the medical costs and *CT* the financial costs of treatment. In Appendix H, we show that allowing for false positives and false negatives results in a model which is isomorphic to the one above with *NU* replaced by $\hat{NU} = \frac{s}{s - fp} MB - MC - CT$ and *c* replaced by $\hat{c} = c + \frac{s \cdot fp}{s - fp} MB$.

Table 6 reports the values of the parameters that we use to compute $\tau * = \frac{\hat{c}}{\hat{NU}}$. Parameters specifying test sensitivity and specificity, the medical benefits of testing, and the medical costs of testing are drawn from the existing medical literature. Note that our calibration of both the medical benefits and the medical cost of treatment depend on an estimate of the value of a statistical life (VSL); following Murphy and Topel (2006) we assume a VSL of $1 million.[23] We estimate the financial cost of testing and the financial cost of PE treatment directly from our Medicare claims data. Appendix Table A.5, which we discuss below, explores the sensitivity of our welfare findings to these calibration parameters.

One parameter of this calibration turns out to be of particular importance and remains a source of uncertainty in the medical literature: the rate of false positive tests. To our knowledge, the single piece of medical evidence on chest CT scans' false positive rate derives from a comparison of CT imaging results to older diagnostic methods, VQ scanning and ultrasonography; the authors estimate the false positive rate at 4% (Stein et al. 2006). We report results with a false positive rate of 4% as our preferred welfare calibration, but also show the welfare implications of assuming a 3% or 0% false positive rate. Lower false positive rates boost the net utility associated with treating a positive test, and thus provide more conservative estimates of the costs of overtesting.

Table 7 reports the optimal testing threshold τ^* under these calibration assumptions. With a false positive rate of 4%, we find physicians should optimally test all patients with an ex ante likelihood of a positive test greater than or equal to 6.2%. The optimal threshold

[23]The choice of a lower VSL estimate in this context is driven by the fact that we are studying an elderly population, with an average age of around 77.

decreases to 5.0% at a false positive rate of 3%; at the (unlikely) extreme of no false positive test results, the optimal threshold falls to 1.5%.

7.2 Welfare impact of eliminating overtesting

The model implies welfare loss whenever a physician's testing threshold τ_d does not equal the optimal value τ^*. We focus on the welfare consequences of overtesting, where τ_d is below this calibrated optimum, for two reasons. First, overtesting is empirically the larger problem in our sample, with an estimated 84% of doctors overtesting under our preferred calibration assumptions. Second, unlike the overtesting case, we find that the welfare loss due to under-testing is highly dependent on the distribution we assume for τ_d when applying an empirical Bayes technique to recover the posterior distribution of τ_d. Previously, we were agnostic about the distribution of τ_d and recovered only the posterior mean and variance, but for welfare calculations, a specific distributional assumption is required. For some distributions of τ_d, even a small number of doctors under-testing can lead to large welfare losses if the right tail of the τ_d distribution is sufficiently thick.

To determine the percentage of doctors overtesting we need to extend our empirical Bayes analysis to recover a posterior estimate of τ_d for each physician; proceeding requires an assumption about the shape of the underlying τ_d distribution. First, note that τ_d is bounded below at the false positive rate. We assume that τ_d minus the false positive rate is log-normally distributed with the posterior mean and variance of the τ_d distribution as previously calculated. Table 7 reports the percentage of doctors overtesting at each false positive rate, given this distributional assumption.

Our initial estimates of τ_d are in units of the probability of a positive test. For example, in our baseline specification, we find that the average doctor tests a patient if the probability of a positive test exceeds 5.6%. We want to know: how would testing behavior change for each physician if all physicians with testing thresholds below $\tau^* = 6.2\%$ instead adopted a threshold of 6.2%? If we observed q_{id} for each patient, this would be a simple matter of counting the number of inframarginal patients. But q_{id} is not observed—instead, we know the probability of a positive test as a function of the propensity to test. Our model allows us to determine how changes in τ_d impact the propensity to test using the scaling factor $\frac{\eta}{p}$, the estimated coefficient on the selection term in equation 14. Equation 14 also allows us to compute how the probability of a positive test conditional on testing changes for each observation. More details are provided in Appendix H.

Combined with our assumptions about costs and net utility, we compute separately the realized medical benefits of testing, the medical costs of testing, the financial costs of testing and the net benefits of testing given the estimated $\hat{\tau}_d$ as well as a counterfactual where $\tau_d = \tau^*$ for all doctors with $\hat{\tau}_d < \tau*$. These results are shown in Table 7, under a series of different assumptions about the false positive rate.

At a false positive rate of 4% (the estimate in the medical literature), we estimate that 84% of the physicians in our sample are overtesting on the margin, i.e. they apply a testing threshold that is lower than the 6.2% threshold probability of a positive test the calibration

suggests is optimal. At a false positive rate of 3%, the proportion of doctors overtesting falls to 67.2%. To illustrate the importance of the false positive rate in assessing welfare, note that if there were no false positive tests, the optimal testing threshold τ^* drops substantially to 1.5% and only 10% of physicians are overtesting on the margin, i.e. have a testing threshold lower than 1.5%.

At a false positive rate of 3% or 4%, eliminating overtesting would decrease the total number of patients tested by more than 30% or 50%, respectively. Why such large effects? Recall that with a false positive rate of 4%, the minimum possible perceived probability (q_{id}) of a positive test is 4%. The median physician in our sample has a τ_d which is less than 5% (much less than the mean, since the distribution is bounded from below by 4%). Increasing τ_d to 6.2% thus greatly increases the range of probabilities q_{id} which would not be tested for many physicians.

In these scenarios, the financial and medical costs of testing would fall by an amount proportional to the decline in tested patients. There would be a small offsetting decline in the medical benefits of testing because the patients not tested in the counterfactual world have a very low probability of truly having a PE. Eliminating overtesting leads to a 12.5% increase in net benefits at a false positive rate of 3% and a more than 60% increase in net benefits at a false positive rate of 4%; the increase in net benefits per test is of course much larger. This exercise illustrates both the large welfare implications of overuse of medical testing and the sensitivity of this result to the false positive rate. As detailed in Table 7, most of the net benefit increase comes from eliminating the financial costs associated with testing low-probability patients for PE and unneeded treatment of patients with false positive test results.

Given the widespread incidence of overtesting under our preferred calibration, it is worth considering a few possible explanations. As we illustrate in Table 7, the estimated overtesting behavior of a majority of doctors in our sample could be explained if they were behaving as if there were no false positive test results. Similarly, if physicians ignored the financial costs associated with testing and treating PE, this could also explain much of the overtesting behavior. However, the only way to rationalize the entire estimated posterior distribution of physician testing patterns would be to allow physicians to vary substantially in their assessment of financial costs or the false positive rate.

One could also interpret variation in τ_d as variation in the patients' "value of knowing" that they do not have a PE. In contrast to the case of Huntington's disease (Oster, Shoulson, and Dorsey 2011), the value of knowing seems an unlikely driver of testing decisions in this context, since in most cases a PE has a very low ex ante probability and the rate of false negatives is sufficiently high that even after testing one has only somewhat reduced that probability. Further, Finkelstein et al. (2014) find that variation in patient demand (i.e. both patient preferences and medical needs) explains only 14% of the regional variation in spending on imaging, suggesting a very limited role for patient preferences in explaining variation in imaging decisions.

Finally, the socially optimal testing threshold depends on the cost of scanning a patient, which we estimate directly from the Medicare claims data. The $300 financial cost of testing

is calculated based on the allowed charges which compensate for the technician's time to run the scan, the radiologist's time to interpret the scan and capital depreciation. If some of this reimbursement is intended as compensation for the high fixed costs of owning a CT scanner, then we may be overstating the social cost of testing. We believe this concern is mitigated by calculating costs directly from the Medicare data, where reimbursement for CT scans remains much below the estimated fees paid by privately insured consumers (cf. Healthcare Blue Book which estimates the typical fee at $517 to $577 depending on the precise billing code). In addition, there may be opportunity costs of scanning a patient not accounted for in our calibration if the hospital is capacity constrained in its allocation of time in the CT scanner or time spent awaiting a scan in an ED bed. If present, opportunity costs would lead us to understate the true costs of performing a scan, and thus understate the amount of overtesting in our data.

Panel A of Table A.5 explores how our results on the net welfare cost of overtesting vary with the calibrated parameters. The results do not vary much with the calibration of test sensitivity. Changing either the VSL or the cost of the test shifts the optimal testing threshold τ^* and thus the welfare benefits. For example, with a VSL of $500,000 rather than $1 million, the optimal threshold increases from 6.2% to 14.3%. Due to this dramatic increase in τ^*, simulations with no physicians overtesting involve more dramatic declines in the fraction of patients tested, and the net benefits of eliminating overtesting almost double vis-a-vis the baseline calibration results. If the VSL is $1.5 million rather than $1 million, the number of patients tested in a world with no overtesting increases by 50%, and the net benefits of eliminating overtesting likewise fall. Similarly, if the cost of the test is $0 (i.e. if there is zero marginal social cost of running a CT scan), the optimal threshold τ^* falls to 4.8%, there is substantially less overtesting and the overtesting that does occur has much lower social cost (only the costs from overtreatment of false positive tests). If the costs of treating patients with positive tests were also equal to $0, the optimal threshold τ^* falls further to 4.3%, eliminating most over-testing but implying large amounts of under-testing. By contrast, if the cost of the test is $500 (comparable to the fees paid to private insurers per CT scan) rather than $300, the net benefits of eliminating overtesting almost double.

7.3 Welfare impact of eliminating misweighting of patient risk factors

Table 8 reports results from a simulation in which doctors select patients for testing by weighting observable comorbidities in the manner the model suggests would maximize detection of positive tests. In other words, we simulate physician behavior if they were to use the true weights β rather than the observed weights β' to assess PE risk. In this simulation, we maintain the distribution of physician testing thresholds at their baseline values, so we allow for the observed patterns of under- and overtesting. We report results at our preferred calibration of the false positive rate, 4%; the welfare consequences of eliminating misweighting would be even larger at lower false positive rates.

Structurally, this exercise is very similar to the exercise where we simulate alternative values of τ_d. Our initial estimates tell us the degree of misweighting in units of the probability of a positive test. We want to determine how the propensity to test would differ if physicians did not misweight; the scaling factor $\frac{\eta}{p}$ allows us to translate the estimated degree of

misweighting into the same units as the testing propensity and calculate the testing propensity and expected test outcomes if there were no misweighting. We demonstrate this explicitly in Appendix H.

One concern with these estimates is that even if there were zero misweighting at the true parameter values, a model like ours would detect some misweighting due to the presence of statistical noise. To deal with this, we conduct a cross-validation exercise where we estimate the scaling factor $\frac{\eta}{p}$ and the misweighting coefficients $\beta - \beta'$ in one half the data (the "training" sample) and then conduct a simulation in the other half (the "test" sample). Test yields are determined by the estimated parameters in the test sample while counterfactual testing decisions are determined by the estimated parameters in the "training" data. These estimates are reported in columns 3 of Table 8. The fact that we find a nearly identical amount of misweighting in the test sample shows that our evaluation of misweighting costs is not driven by statistical noise.

We find that properly weighting observables to improve PE detection would lead the fraction of patients tested to increase from 3.8% to 4.3%, by moving some patients just over their estimated physician's testing threshold. But by far the predominant welfare impact comes from the predicted increase in the rate of PE detection. The medical benefits due to treatment of PE nearly double and the net benefits of testing more than triple. The total welfare loss from misweighting ($35.9 million in our sample) is more than 4 times as large as the welfare loss from overtesting ($8.1 million), even in the model with the highest rate of false positives.

To investigate whether a small number of risk factors account for most of the observed costs of misweighting, we conduct an exercise where we correct the weights applied to each variable, one at a time. Results from this exercise with more detailed notes are reported in Appendix Table A.4. First, it is worth noting that in this simulated second-best world where physicians do not all share the optimal testing threshold τ^* and where other factors are misweighted, correcting misweighting of a single risk factor in isolation can sometimes worsen total welfare; certain misweighting errors offset some of the costs associated with overtesting. However, in most cases, correcting a single variable's weight weakly improves estimated welfare.

Correcting the weighting on 30-day inpatient admissions accounts for approximately 20% of the total potential gains from eliminating misweighting. Expanding the list to include the 5 highest-impact covariates (30-day admission history, 1-week admission history, 1-year surgical history, chronic obstructive pulmonary disease, and ischemic heart disease) accounts for roughly 60% of the total potential gains. These covariates are both substantially misweighted and common enough to induce large welfare consequences.

Intuitively, given our estimates of misweighting in Section 5.2, it is not surprising that the welfare loss from misweighting substantially exceeds the welfare losses from overtesting. Several factors combine to make misweighting a more serious problem. Physicians behave as if they are misestimating a patient's PE risk by 2.3 percentage points on average by failing to weight observable characteristics to maximize detection of positive tests. By

comparison, the average difference between τ_d and τ^* for physicians who are overtesting is only 1.7 percentage points in the calibration with a false positive rate of 4%. The welfare cost of misweighting errors or suboptimal values of τ_d increases with the square of the deviation—as the bias grows, both the number of patients impacted and the average severity of the error among those patients increases. Further, the welfare costs of overtesting are bounded. The worst outcome of overtesting is that a patient is tested with no chance of having a PE and incurs the cost of the test (a few hundred dollars) plus the potential financial costs and medical risk of treatment if they receive a false positive test result. The potential costs of misweighting are substantially greater since you might fail to treat a patient with a substantial risk of death.

Panel B of Table A.5 explores how our results on the net welfare cost of misweighting vary with the calibrated parameters. The positive impact of misweighting on testing behavior does not depend on the calibration (unlike the case of overtesting, since the calibration determines which physicians overtest). The welfare cost of misweighting is not too sensitive to the false positive rate, the sensitivity of the test or the cost of the test, but it is sensitive to the VSL. Misweighting creates more welfare loss from undertesting than overtesting: the welfare costs of overtesting are bounded by the financial costs of the test plus the costs of treating false positive test results, while the costs of undertesting in the worst case is the 2.5% chance of mortality from a missed PE. These latter costs are roughly proportional to the VSL.

Undiagnosed PE is thought to be a major public health problem, with the Office of the Surgeon General (2008) estimating that approximately half of PE cases are never diagnosed; analysis of autopsy reports have found it to be a frequently missed mortality risk. By improving physician assessment of patient PE risk, our model suggests that the rate of undiagnosed PE could fall substantially. Although there is policy attention in the medical community on the risks associated with the perceived overuse of PE CT, this evidence suggests that there may be even larger gains possible from improving the targeting of CT scans.

Our welfare calculations are based on a 20% sample of patients enrolled in Medicare Parts A and B over a 10-year period, and the numbers reported in Tables 7 and 8 reflect potential gains to this sample only. To understand the annual welfare loss for Medicare patients associated with the inefficiencies we identify in this sample, we do an informal scaling exercise. We first scale the estimates up by a factor of 5 to account for the entire population of Medicare fee for service enrollees, then adjust to account for the 28% of Medicare patients who enroll in a Medicare Advantage plan, and finally divide by 10 to calculate annual estimates. We recover a $5.5 million annual welfare loss from overuse of PE CT due to low testing thresholds, and a $25 million annual loss from misweighting observable patient risk factors, for emergency department CT scans among elderly patients. Yet these scaled welfare gains from the efficient application of PE CT to the elderly population seeking emergency care may represent only a small fraction of the total welfare benefit available from more efficient diagnostic testing and treatment decisions across a variety of medical conditions.

8 Conclusion

While it is commonly believed that the US health care system spends significant resources on services that have low medical returns and high costs, there is little consensus on how this waste could be reduced. Wasteful spending is characterized both by overuse of medical care and mistargeting of medical resources. This paper investigates both forms of inefficiency, analyzing whether doctors efficiently select patients for medical testing and how physicians vary in the risk thresholds at which they test patients. We study these inefficiencies in the context of emergency department CT scans to diagnose pulmonary embolism (PE). We document both widespread variation in physician use of CT scans for PE unexplained by differences in patient risk, and also systemic failure to target medical testing to the highest risk patients.

The identification strategy underlying this analysis relies on exclusion restrictions motivated by our structural model of testing behavior. The identification arguments require that physicians select patients for testing on the basis of private information about expected PE risk and they apply a consistent PE risk threshold across patients. The ignorability assumption that private information about PE risk is independently and identically distributed across doctors and patients underlies the single-index structure and is important to identifying the model. Further, to the extent that we have not isolated marginal tested patients to recover test thresholds of high-volume doctors, we may be understating the full costs of over-testing behavior; notably, sensitivity analyses suggest our results on misweighting are not sensitive to the definition of marginal patients. If the value of treating pulmonary embolism varies substantially across patients, this may explain some of the apparent patterns of misweighting and overuse.

Estimating the model to study physicians' CT scanning decisions in a national sample of Medicare claims, we find substantial variation in physician's use of diagnostic scans on low-risk patients. This variation generates a negative relationship between testing propensities and test yield across physicians, since physicians who test more also test lower risk patients on average. Investigating the role of training and practice environment in explaining practice styles, we find that physicians practicing in high-spending Dartmouth Atlas regions and those with less experience are more likely to scan low-risk patients. Other factors, such as hospital ownership or quality of medical school training are not significantly related to testing behavior. Taken as a whole, observable characteristics can explain only a small fraction of the total variation in testing thresholds. Applying further calibration assumptions suggests that 84% of physicians in our sample are overtesting on the margin in the sense that their risk threshold is lower than the calibrated optimum.

We also find that doctors do not weight observable patient risk factors in a way that would maximize test yields. Physicians systematically underweight certain important predictors of PE risk, including recent prior hospitalizations and metastatic cancer. Other preexisting conditions that have similar clinical symptoms to PE are over-weighted in the testing decision. These apparent errors occur despite the fact that physicians are widely encouraged to use diagnostic scoring systems such as the Wells or Geneva score to assess the risk of PE before deciding whether to order a CT scan. The continued prevalence of risk assessment

mistakes despite the popularity of these PE risk scoring systems may reflect shortcomings in the scoring systems themselves or failures to make adequate use of these scores. (The data used in this project cannot disentangle these possibilities.) Together, these mistakes in assessing patient PE risk lead to significant welfare losses from failing to target the test to the highest risk patients according to our welfare simulations. In fact, despite the huge attention in the health economics literature to the problem of overuse of care, the simulated welfare loss from mistargeting of diagnostic imaging is four times larger than the welfare loss from overuse.

The model developed in this paper could be applied to a variety of empirical contexts—it is applicable whenever economic actors make repeated decisions about whom to treat, as long as the objective function is known for the counterfactual where treated individuals are untreated. In the PE testing case, we know that untested individuals have no PE detected. In other applications, the model could be used to evaluate the decisions of loan officers to extend credit, hiring directors to select among potential job applicants, or admissions officers to predict which students will perform most highly. Positively, one could investigate the degree to which observed heterogeneity in treatment rates is due to decision-maker discretion. Normatively, many of these organizations have specific objectives they seek to optimize (e.g. reducing default on loans or productivity among employees) and one could use the model developed here to investigate whether observed selection patterns are successfully optimizing these outcomes.

Our findings suggest that both overuse and misuse of medical resources are important drivers of high spending and low medical returns to care. Future work could pair this framework for estimating overuse of diagnostic testing with experimental or quasi-experimental variation in physician's training or practice environment; these estimates could more directly inform policy by causally identifying how these changes to a physician's education or training affect the efficiency of the medical care delivered. Given more detailed patient-level data, our model could be used to formulate optimal guidelines and risk scores, overcoming the selection problems that may lead to biased estimates of risk under popular existing methodologies. Our findings underscore the fact that purely cost-focused health reform may be insufficient to achieve efficiency in healthcare delivery—there are potentially large benefits to patients from physicians making better use of the available information to target medical resources to those patients with the highest returns.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Abaluck J, Agha L, Chan D. Discretion and Guidelines, Evidence from Warfarin Administration. Working Paper. 2016

Altonji JG, Elder TE, Taber CR. Using selection on observed variables to assess bias from unobservables when evaluating swan-ganz catheterization. The American Economic Review. 2008; 98(2):345–350.

Avraham R. Database of state tort law reforms (dstlr 4th). U of Texas Law, Law and Econ Research Paper. 2011; (184)

Avraham R, Dafny LS, Schanzenbach MM. The impact of tort reform on employer-sponsored health insurance premiums. Journal of Law, Economics, and Organization. 2012; 28(4):657–686.

Chandra A, Staiger D. Expertise, Overuse and Underuse in Healthcare. Working Paper. 2011

Chandra, A., Staiger, DO. Working Paper 16382. National Bureau of Economic Research; 2010 Sep. Identifying provider prejudice in healthcare.

Coco AS, O'Gurek DT. Increased emergency department computed tomography use for common chest symptoms without clear patient benefits. Journal of the American Board of Family Medicine. 2012 Jan-Feb;25(1):33–41. [PubMed: 22218622]

Costantino MM, Randall G, Gosselin M, Brandt M, Spinning K, Vegas CD. Ct angiography in the evaluation of acute pulmonary embolus. American Journal of Roentgenology. 2008 Aug; 191(2): 471–474. [PubMed: 18647919]

Currie, J., MacLeod, WB. Technical report. National Bureau of Economic Research; 2006. First do no harm?: Tort reform and birth outcomes.

Currie, J., MacLeod, WB. Technical report. National Bureau of Economic Research; 2013. Diagnosis and unnecessary procedure use: Evidence from c-section.

Cutler, D., Skinner, J., Stern, AD., Wennberg, D. Technical report. National Bureau of Economic Research; 2013. Physician beliefs and patient preferences: a new look at regional variation in health care spending.

David S, Beddy P, Babar J, Devaraj A. Evolution of ct pulmonary angiography: referral patterns and diagnostic yield in 2009 compared with 2006. Acta Radiologica. 2012 Feb; 53(1):36–43.

Doyle JJ, Ewer SM, Wagner TH. Returns to physician human capital: Evidence from patients randomized to physician teams. Journal of health economics. 2010; 29(6):866–882. [PubMed: 20869783]

Elixhauser A, Steiner C, Harris D, Coffey R. Comorbidity measures for use with administrative data. Medical Care. 1998; 36(1):8–27. [PubMed: 9431328]

Finkelstein, A., Gentzkow, M., Williams, H. Technical report. National Bureau of Economic Research; 2014. Sources of geographic variation in health care: Evidence from patient migration.

Garber, AM., Skinner, J. Technical report. National Bureau of Economic Research; 2008. Is american health care uniquely inefficient?.

Goldhaber SZ, Bounameaux H. Pulmonary embolism and deep vein thrombosis. The Lancet. 2012; 379(9828):1835–1846.

Heckman J, MaCurdy T. A life cycle model of female labour supply. The Review of Economic Studies. 1980; 47(1):47–74.

Heckman JJ. Sample selection bias as a specification error. Econometrica. 1979; 47(1):153–161.

Heckman JJ, Vytlacil E. Structural equations, treatment effects, and econometric policy evaluation1. Econometrica. 2005; 73(3):669–738.

Iglehart JK. Health insurers and medical-imaging policya work in progress. New England Journal of Medicine. 2009; 360(10):1030–1037. [PubMed: 19264694]

Jackson CK, Rockoff JE, Staiger DO. Teacher effects and teacher-related policies. Annual Review of Economics. 2014; 6(1):801–825.

Kane, TJ., Staiger, DO. Technical report. National Bureau of Economic Research; 2008. Estimating teacher impacts on student achievement: An experimental evaluation.

Klein R, Spady R. An efficient semiparametric estimator for binary response models. Econometrica: Journal of the Econometric Society. 1993:387–421.

Lessler AL, Isserman JA, Agarwal R, Palevsky HI, Pines JM. Testing low-risk patients for suspected pulmonary embolism: A decision analysis. Annals of Emergency Medicine. 2010 Apr; 55(4):316–326. [PubMed: 20061065]

Lewis JB, Linzer DA. Estimating regression models in which the dependent variable is based on estimates. Political Analysis. 2005; 13(4):345–364.

Mamlouk MD, vanSonnenberg E, Gosalia R, Drachman D, Gridley D, Zamora JG, Casola G, Ornstein S. Pulmonary embolism at ct angiography: Implications for appropriateness, cost, and radiation exposure in 2003 patients. Radiology. 2010 Aug.256:625–632. [PubMed: 20551182]

Meszaros I, Morocz J, Szlavi J, Schmidt J, Tornoci L, Nagy L, Szep L. Epidemiology and clinicopathology of aortic dissection. Chest. 2000 May; 117(5):1271–1278. [PubMed: 10807810]

Molitor D. The evolution of physician practice styles evidence from cardiologist migration. Technical report, MIT working paper. 2012

Mulligan CB, Rubinstein Y. Selection, investment, and women's relative wages over time. The Quarterly Journal of Economics. 2008; 123(3):1061–1110.

Murphy KM, Topel RH. The value of health and longevity. Journal of Political Economy. 2006; 114(5):871–904.

Nelson A. Unequal treatment: confronting racial and ethnic disparities in health care. Journal of the National Medical Association. 2002; 94(8):666. [PubMed: 12152921]

Office of the Surgeon General. The surgeon general's call to action to prevent deep vein thrombosis and pulmonary embolism. 2008

Oster, E., Shoulson, I., Dorsey, E. Technical report. National Bureau of Economic Research; 2011. Optimal expectations and limited medical testing: evidence from huntington disease.

Rahimtoola A, Bergin JD. Acute pulmonary embolism: An update on diagnosis and management. Current Problems in Cardiology. 2005 Feb.30:61–114. [PubMed: 15650680]

Rao VM, Levin DC. The overuse of diagnostic imaging and the choosing wisely initiative. Annals of internal medicine. 2012; 157(8):574–576. [PubMed: 22928172]

Stein PD, Fowler SE, Goodman LR, Gottschalk A, Hales CA, Hull RD, Kenneth J, Leeper V, John Popovich J, Quinn DA, Sos TA, Sostman HD, Tapson VF, Wakefield TW, Weg JG, Woodard PK. Multidetector computed tomography for acute pulmonary embolism. New England Journal of Medicine. 2006 Jun 1; 354(22):2317–27. [PubMed: 16738268]

Venkatesh A, Kline JA, Kabrhel C. Computed tomography in the emergency department setting-reply. Journal of the American Medical Association Internal Medicine. 2013 Jan 28; 173(2):167–168. [PubMed: 23358840]

Venkatesh AK, Kline JA, Courtney DM, C CA Jr, Plewa MC, Nordenholz KE, Moore CL, Richman PB, Smithline HA, Beam DM, Kabrhel C. Evaluation of pulmonary embolism in the emergency department and consistency with a national quality measure: Quantifying the opportunity for improvement. Archives of Internal Medicine. 2012 Jul 9; 172(13):1028–1032. [PubMed: 22664742]

Wells PS, Anderson DR, Rodger M, Ginsberg JS, Kearon C, Gent M, Turpie A, Bormanis J, Weitz J, Chamberlain M, Bowie D, Barnes D, Hirsh J. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism-increasing the models utility with the simplired d-dimer. Thrombosis and Haemostasis. 2000; 83(3):416–420. [PubMed: 10744147]

Wells PS, Ginsberg JS, Anderson DR, Kearon C, Gent M, Turpie AG, Bormanis J, Weitz J, Chamberlain M, Bowie D, Barnes D, Hirsh J. Use of a clinical model for safe management of patients with suspected pulmonary embolism. Annals of internal medicine. 1998; 129(12):997–1005. [PubMed: 9867786]

Wells PS, Hirsh J, Anderson DR, Lensing AWA, Foster G, Kearon C, Weitz J, D'Ovidio R, Cogo A, Prandoni P, Girolami A, Ginsberg JS. Accuracy of clinical assessment of deep-vein thrombosis. The Lancet. 1995; 345(8961):1326–1330.

Wennberg, J., Cooper, M., et al. The Dartmouth atlas of health care in the United States. Chicago, IL: American Hospital Association; 1996.

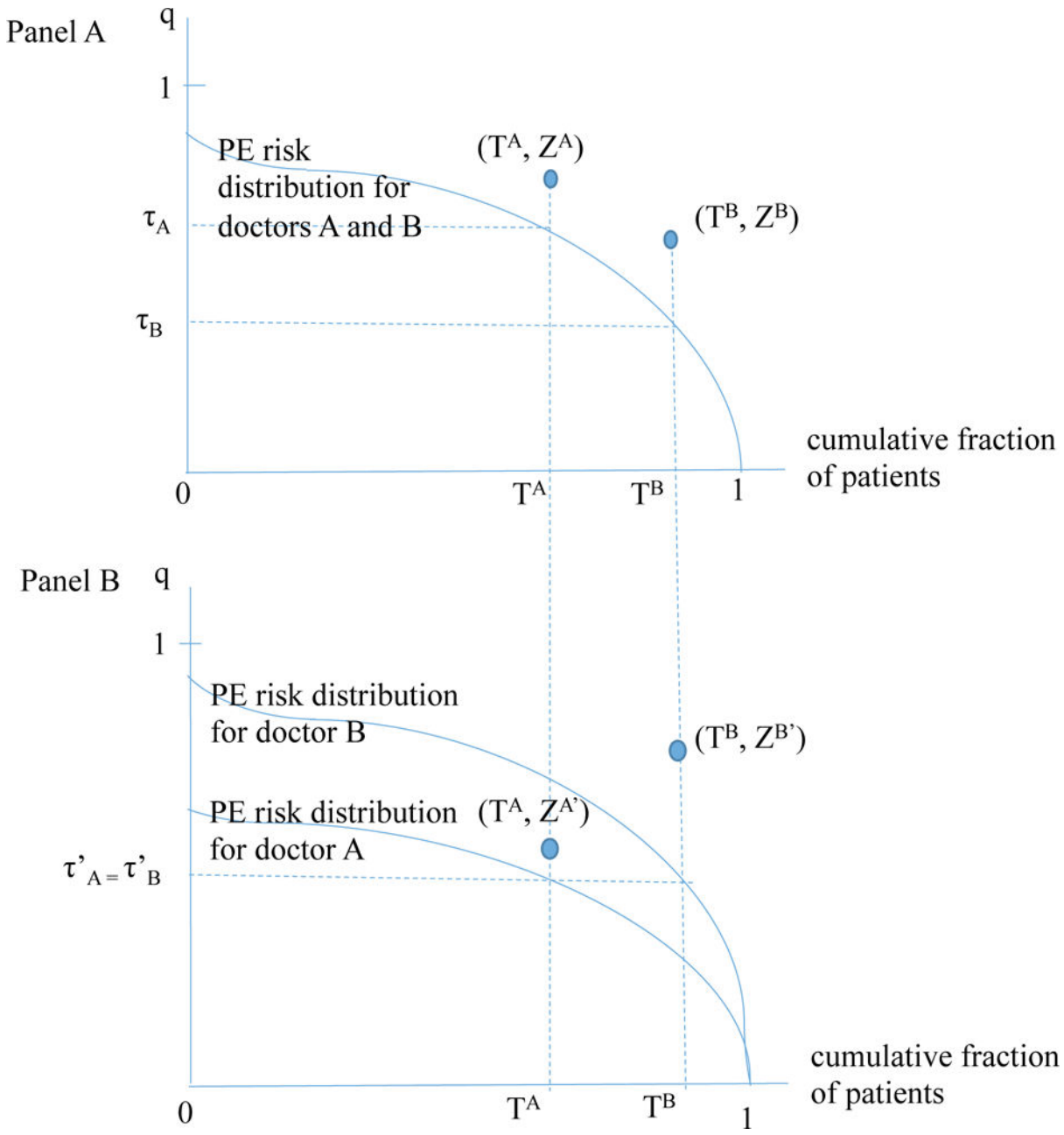

Figure 1.

Stylized relationship between testing thresholds, testing rates, and test yields

Notes: Figure illustrates the theoretic relationship between testing thresholds, test yields and fraction of patients tested for two hypothetical doctors, A and B. Patients are sorted along the x-axis according to their risk of PE, q_{id}, from highest risk to lowest risk. Each point (x, y) along the plotted curve shows the fraction of patients x for whom $q_{id} \geq y$. For example, at point ($T^A = 2/3$, $\tau^A = 1/2$) in Panel A, the graph indicates that 2/3 of patients have a risk of PE that equals or exceeds 1/2. τ_A denotes doctor A's testing threshold, T^A denotes the fraction of patients tested by doctor A, Z^A denotes doctor A's test yield (among tested patients), and likewise for doctor B. In Panel A, both doctors face patient populations with

the same distribution of PE risk. In Panel B, Doctor B's patients are higher risk, i.e. for any given probability of a positive test q, a greater fraction of doctor B's patients meet or exceed that threshold compared to doctor A.

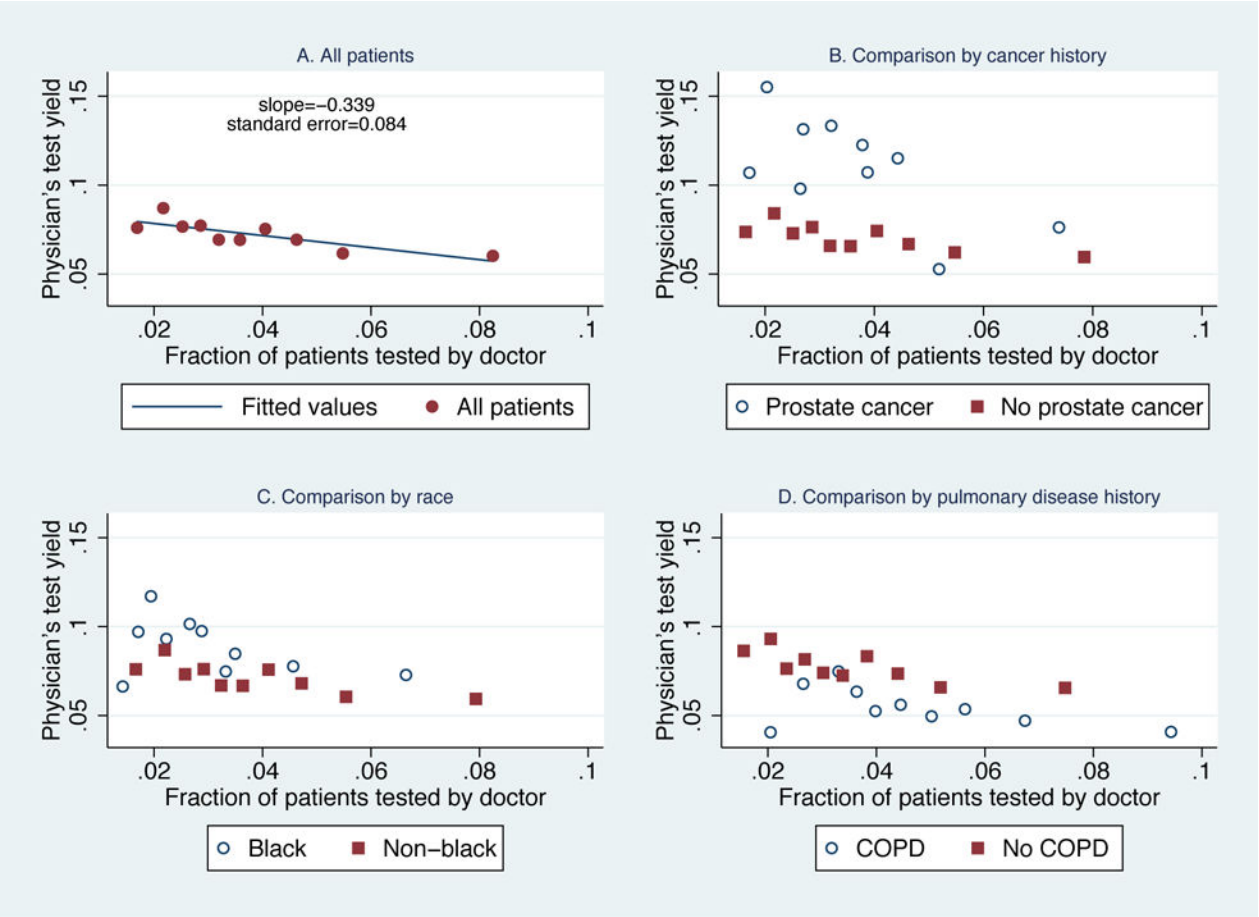

Figure 2.

Binned scatterplot of physician test yield by fraction of patients tested

Notes: Figures displays a binned scatterplot based on our sample of Medicare claims data. Physicians are binned into deciles according to the fraction of patients they test. Panel A reports results across all patients evaluated by each doctor; the x-axis reports the average fraction of patients tested and the y-axis reports the rate of positive test results among tested patients, within each physician decile. The slope coefficient and standard error on the simple bivariate regression of average test yield on fraction of patients tested is reported on the panel. Panels B, C, and D maintain the same definitions of physician groups by deciles of test rate as in Panel A, but splits each doctor's patients into groups according to whether they have a particular risk characteristic. We report average test rates and test yields by physician's test decile, for patients with and without the listed characteristic.

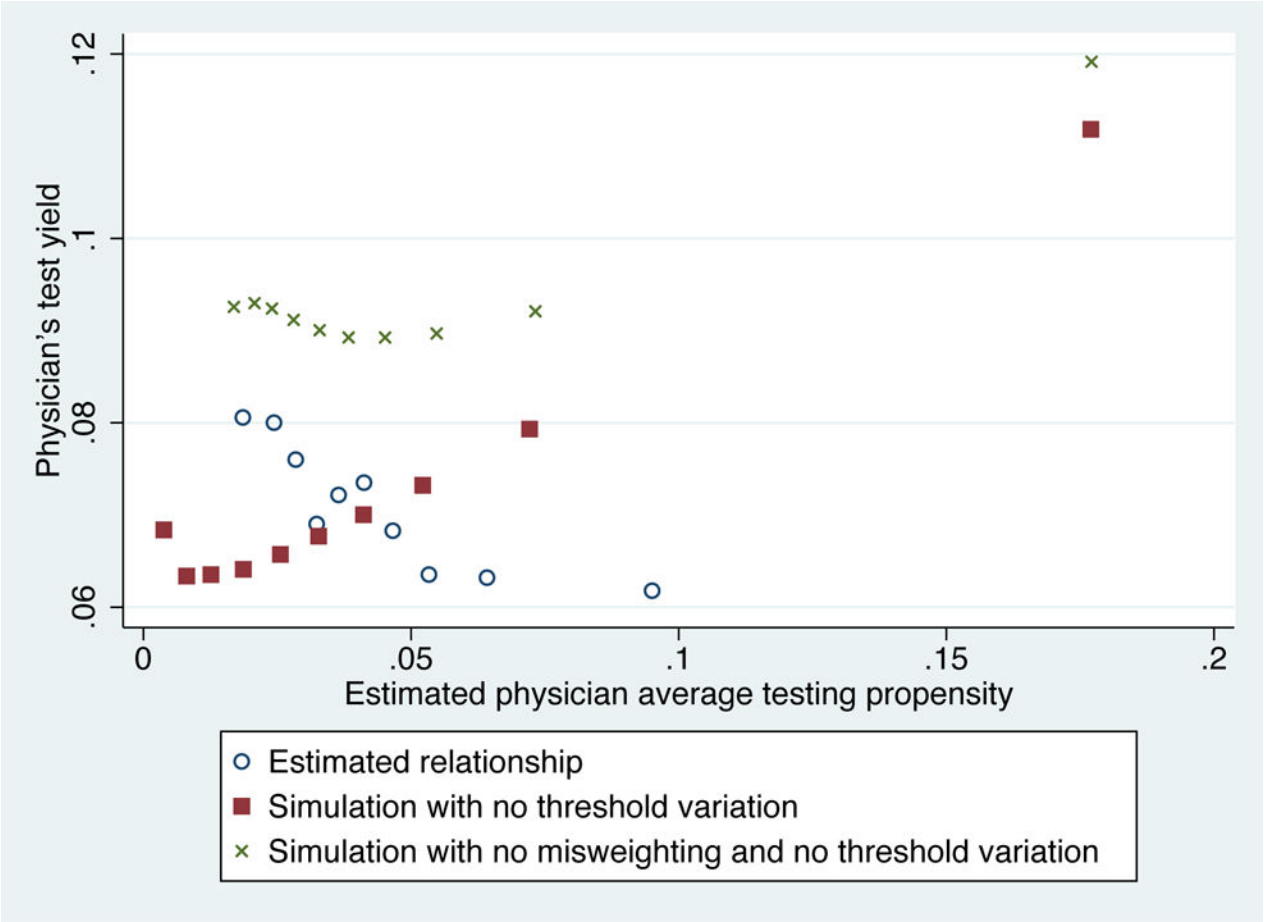

Figure 3.

Binned scatterplot of physician test yield by testing propensity index: Estimation results and simulations

Notes: Figure displays a binned scatterplot based on our estimation and simulation results; physicians are binned into deciles based on the average estimated value of the testing propensity index $\tilde{I}'_{id}$. The open circle markers plots the relationship between physicians' actual test yields and physicians' average $\tilde{I}'_{id}$. Th solid square markers display the simulated relationship between testing propensities and test yields under a counterfactual with no variation in physician testing thresholds, and instead all physicians assigned the average testing threshold $E(\tau_d)$. The X-shaped markers displays the simulated relationship between testing propensities and test yields if there were no variation in physician testing thresholds *and* there were no misweighting of observable risk factors.

Table 1

Summary statistics

	A. Untested patients	*B. Patients with negative tests*	*C. Patients with positive tests*
Patient characteristics			
Age	77.6	76.8	76.9
Female	0.586	0.602	0.600
Black	0.082	0.066	0.083
History of PE	0.003	0.006	0.017
Doctor, hospital and region characteristics			
Doctor experience	16.5 (8.3)	16.4 (8.4)	16.8 (8.5)
Top 50 research med. school	0.28	0.29	0.30
Top 50 primary med. school	0.26	0.27	0.28
Academic hospital	0.33	0.34	0.356
For profit hospital	0.12	0.13	0.120
HRR avg spending (in $)	8,198 (959)	8,173 (972)	8,089 (936)
Average income in region	22,771 (5521)	23,005 (5490)	23,039 (5710)
Joint and several liability	0.69	0.70	0.692
Malpractice damage caps	0.70	0.76	0.747
Number of observations	1,819,015	66,677	4,968

Notes: Table reports means and standard deviations (in parentheses). Data is from the Medicare claims 2000–2009, the American Hospital Association annual survey, the American Medical Association Masterfile, the Dartmouth Atlas, and the Avraham Database of State Tort Law Reform.

Table 2

Summary statistics illustrating potential misweighting of risk factors

	A. Fraction tested	B. Test yield
Selected candidates for under-weighting		
Prostate cancer (CCW)	0.0370	0.1019
No prostate cancer (CCW)	0.0380	0.0677
Black	0.0313	0.0851
Non-black	0.0385	0.0682
History of PE	0.0726	0.1881
No history of PE	0.0378	0.0686
History of deep vein thrombosis	0.0507	0.1656
No history of deep vein thrombosis	0.0378	0.0685
Prior hospital visit within 30 days	0.0465	0.1976
No prior hospital visit within 30 days	0.0377	0.0656
Selected candidates for over-weighting		
Chronic obstructive pulmonary disease (CCW)	0.0466	0.0524
No chronic obstructive pulmonary disease (CCW)	0.0360	0.0742
Ischemic heart disease (CCW)	0.0376	0.0566
No ischemic heart disease (CCW)	0.0382	0.0786
Atrial fibrillation (CCW)	0.0317	0.0520
No atrial fibrillation (CCW)	0.0388	0.0713

Notes: Table reports summary statistics for selected comorbidities to motivate the examination of misweighting. Variables are selected on the Column A reports average rates of testing for patients with and without the listed conditions. Column B reports average rate of positive tests among tested patients with and without the listed conditions. CCW notes comorbidity is coded by the Chronic Condition Warehouse. Data is from the Medicare claims 2000–2009.

Table 3

Regressions of testing threshold on physician characteristics and practice environment

	Dependent variable: Physician testing threshold τ_d			
	OLS	**FGLS**	**OLS**	**FGLS**
Independent variables:	(1)	(2)	(3)	(4)
Doctor experience	0.0007*** (0.0001)	0.0007*** (0.0001)	0.0007*** (0.0002)	0.0008*** (0.0001)
Top 50 research medical school	0.0047 (0.0038)	0.0050 (0.0031)	0.0053 (0.0047)	0.0032 (0.0037)
Top 50 primary care medical school	−0.0062 (0.0039)	−0.0042 (0.0032)	−0.0077 (0.0048)	−0.0030 (0.0037)
Academic hospital	0.0006 (0.0026)	0.0007 (0.0022)		
For profit hospital	−0.0004 (0.0041)	−0.0018 (0.0032)		
Log(HRR average Medicare spending)	−0.0391*** (0.0109)	−0.0474*** (0.0093)		
Average income in region (in $10k)	0.0000 (0.0025)	0.0000 (0.0019)		
Joint and several liability	0.0001 (0.0027)	0.0003 (0.0023)		
Malpractice damage caps	−0.0029 (0.0028)	−0.0053** (0.0023)		
Hospital Fixed Effects	No	No	Yes	Yes

Notes: Each column reports results from a regression of estimated physician testing thresholds τ_d on characteristics of the physician's training and practice environment. Even numbered columns report FGLS estimates which account for estimation error in τ_d. Columns 3 and 4 include hospital fixed effects. An observation is an individual doctor; there are 6636 observations.

* significant at the 10% level

** significance at the 5% level;

*** significance at the 1% level.

Table 4

Comorbidities with significant misweighting: Impact of comorbidity on testing decisions and estimated misassessment of PE risk

	Marginal effect from testing eqn (1)	Misassessment of PE risk (2)	Std. error of misassessment (3)	T statistic of misassessment (4)
Underweighted risk factors				
Prior hospital visit w/in 30 days	−0.0094	0.1070	0.0121	8.8430
Prior hospital visit w/in 7 days	−0.0041	0.1128	0.0130	8.6769
Prostate cancer (CCW)	0.0014	0.0298	0.0048	6.2083
Cancer metastisis (Elixhauser)	−0.0155	0.0726	0.0128	5.6719
History of deep vein thrombosis	0.0092	0.0571	0.0114	5.0088
History of pulmonary embolism	0.0315	0.0666	0.0145	4.5931
Rhumatoid arthritis, osteoarthritis (CCW)	0.0053	0.0091	0.0024	3.7917
Endometrial cancer (CCW)	−0.0011	0.0547	0.0153	3.5752
Obesity (Elixhauser)	0.0095	0.0218	0.0076	2.8684
Paralysis (Elixhauser)	−0.0026	0.0331	0.0117	2.8291
Other neurological conditions (Elixhauser)	−0.0043	0.0194	0.0075	2.5867
Any prior admission history	0.0028	0.0102	0.0041	2.4878
Alzheimer's disease (CCW)	−0.0023	0.0152	0.0064	2.3750
Colorectal cancer (CCW)	−0.0012	0.0136	0.0067	2.0299
Overweighted risk factors				
Ischemic heart disease (CCW)	0.0007	−0.0226	0.0023	−9.8261
Chronic obstructive pulmonary disease (CCW)	0.0132	−0.0182	0.0036	−5.0556
Atrial fibrillation (CCW)	−0.0066	−0.0156	0.0036	−4.3333
Depression (Elixhauser)	0.0033	−0.0208	0.0069	−3.0145
Peripheral vascular disease (Elixhauser)	−0.0013	−0.0214	0.0071	−3.0141
Diabetes (CCW)	−0.0055	−0.0087	0.0029	−3.0000
Osteoperosis (CCW)	0.0024	−0.0087	0.0033	−2.6364
Deficiency anemias (Elixhauser)	−0.0004	−0.0142	0.0056	−2.5357
Asthma (CCW)	0.0043	−0.0088	0.0040	−2.2000
Chronic pulmonary disease (Elixhauser)	−0.0042	−0.0094	0.0048	−1.9583
Demographic factors				
Black	−0.0074	0.0257	0.0044	5.8409
Asian	0.0005	−0.0386	0.0118	−3.2712
Hispanic	−0.0056	−0.0168	0.0097	−1.7320
Female	0.0014	0.0000	0.0024	0.0000
Age 65–69	−0.0012	0.0119	0.0037	3.2162
Age 70–74	−0.0089	0.0129	0.0052	2.4808
Age 75–79	−0.0024	0.0140	0.0038	3.6842
Age 80–84	−0.0033	0.0166	0.0039	4.2564
Age 85–89	−0.0043	0.0208	0.0042	4.9524
Age 90–94	−0.0127	0.0132	0.0078	1.6923

Notes: This table reports results only for demographic variables and variables with statistically significant evidence of misweighting. The results are continued in Appendix Table A.2, which reports results for the remaining comorbidities. Column 1 reports marginal effects from coefficient estimates of the testing equation (i.e. equation 2); for example, patients who were admitted to the hospital within 30 days are 0.94 percentage points less likely to be tested, after controlling for included PE risk factors and physicians' testing thresholds. Column 2 reports estimates of physicians' misweighting of these PE risk factors estimated from equation 14; for example, physicians' observed testing patterns suggest they are underestimating the PE risk associated with a prior hospital visit in the past 30 days by 10.7 percentage points. Column 3 reports standard errors on these misweighting terms. Column 4 reports t-statistics. Variables are sorted by statistical significance, with the exception of demographic risk factors.

Table 5

Distribution of testing thresholds and misweighting under alternative estimation strategies

	Baseline parametric model, all comorbidities	Parametric model, Elixhauser comorbidities excluded	Parametric model, Elixhauser comorbidities & demographics excluded
	(1)	(2)	(3)
Mean of τ_d	0.0563	0.0623	0.0662
Standard deviation of τ_d	0.0540	0.0396	0.0394
Average absolute value of PE misassessment	0.0226	0.0214	0.0200
Standard deviation of PE misassessment	0.0347	0.0336	0.0329
Number of observations	1,890,660	1,890,660	1,890,660
	Heteroskedastic parametric model	Semiparametric model, linear polynomial	Semiparametric model, cubic polynomial
	(4)	(5)	(6)
Mean of τ_d	0.0703	0.0672	0.0661
Standard Deviation of τ_d	0.0514	0.0539	0.0541
Average absolute value of PE misassessment	0.0212	0.0207	0.0208
Standard deviation of PE misassessment	0.0361	0.0357	0.0364
Number of observations	861,707	861,707	861,707

Notes: Panel 1 reports the estimated posterior mean and standard deviation of physician testing thresholds τ_d from our baseline parametric model, after applying the Bayesian shrinkage described in Appendix F. Recall that τ_d is the threshold probability of a positive test at which a physician determines it is worthwhile to test a patient. The average absolute value of misweighting calculates the absolute value of the difference between physicians' assessment of the patient's PE probability and the estimated risk associated with the patient's comorbidities, and then averages this value across all patients. The standard deviation of misweighting describes how the amount of misweighting varies across patients. Panel 2 reports results from the parametric model that excludes all Elixhauser comorbidities. Panel 3 reports results from the parametric model that excludes both Elixhauser comorbidities and demographic variables. Panel 4 reports results from the heteroskedastic model described in Section 6.2, which allows the variance of η_{id} to differ across physicians. Panels 5 and 6 report results from the semiparametric model described in Section 6.3, where Panel 5 fits the function $\lambda(\cdot)$ with a linear function and Panel 6 applies a cubic polynomial. Models estimated in Panels 4, 5, and 6 exclude Elixhauser comorbidities and demographic variables and are estimated on a random subsample of half of the physicians for computational tractability.

Table 6

Calibration Parameters

Definition	Value	Parameter	Source
test sensitivity	0.83	s	Stein et al., 2006
baseline false positive rate	0.04	fp	Stein et al., 2006
value of a statistical life	$1,000,000	VSL	Murphy and Topel, 2006
medical benefit of treating PE	0.025 VSL	MB	Lessler et al., 2009
medical cost of treating PE	0.0017 VSL	MC	Lessler et al., 2009
financial cost of testing	$300	c	estimated from Medicare claims
financial cost of PE treatment	$2,800	CT	estimated from Medicare claims

Notes: Calibrated parameters of the model applied in welfare simulations reported in Section 7.

Table 7

Patient welfare with observed testing thresholds vs. in simulations with no overtesting

	False positive rate of 4 percent $\tau^*=0.062$		*False positive rate of 3 percent* $\tau^*=0.050$		*False positive rate of 0 percent* $\tau^*=0.015$	
	Actual (1)	**Simulation (2)**	**Actual (3)**	**Simulation (4)**	**Actual (5)**	**Simulation (6)**
Description of simulation results:						
Fraction of doctors over-testing	83.7%	0%	67.2%	0%	10.4%	0%
Percent of patients tested	3.8%	1.9%	3.8%	2.6%	3.8%	3.7%
Number of patients tested	71,314	35,140	71,314	49,390	71,314	70,497
Test yield among tested patients	7.0%	9.0%	7.0%	8.3%	7.0%	7.1%
Welfare analysis:						
Total financial costs of testing ($ millions)	35.6	19.5	35.6	26.4	35.6	35.3
Total medical cost of testing ($ millions)	8.5	5.4	8.5	6.9	8.5	8.5
Total medical benefits of testing ($ millions)	57.5	46.3	74.6	67.6	125.0	124.8
Net benefits of testing ($ millions)	13.5	21.4	30.4	34.2	80.9	81.0
Total (financial + medical) costs per test ($)	618.9	709.1	618.9	675.3	618.9	621.2
Total benefits per test ($)	806.9	1318.7	1045.5	1368.3	1752.8	1770.5
Net benefits per test ($)	188.1	609.6	426.7	693.0	1134.0	1149.3

Notes: We compare testing behavior and social welfare under the estimated posterior distribution of physician testing thresholds τ_d (in odd numbered columns) to simulated behavior assuming all physicians with thresholds below the calibrated optimum are reassigned to the optimal testing threshold of $\tau_d = \tau^*$ (in even numbered columns). The simulated results do not correct for misweighting. We report results under three different assumptions about the rate of false positive test results, described in the column headers.

Am Econ Rev. Author manuscript; available in PMC 2017 November 01.

Table 8

Patient welfare with observed misweighting vs. in simulations with no misweighting

	False positive rate of 4%		
	Actual testing decisions	**No misweighting, simulation without cross validation**	**No misweighting, simulation with cross validation**
	(1)	**(2)**	**(3)**
Description of results:			
Percent of patients tested	3.8%	4.3%	4.3%
Number of patients tested	71314	81410	79734
Test yield among tested patients	7.0%	9.2%	8.6%
Number of positive tests detected	5019	7526	6872
Welfare analysis:			
Total financial costs of testing ($ millions)	35.6	45.2	43.4
Total medical cost of testing ($ millions)	8.5	12.4	11.7
Total medical benefits of testing ($ millions)	57.5	106.8	96.7
Net benefits of testing ($ millions)	13.5	49.1	41.6
Total (financial + medical) costs per test ($)	618.9	707.8	690.8
Total benefits per test ($)	806.9	1311.3	1213.1
Net benefits per test ($)	188.1	603.5	522.2

Notes: We compare testing behavior and social welfare under the observed physician weighting of patient risk factors (in column 1) to simulated behavior assuming that physicians target testing to patients with the highest expected probability of a positive test based on observable demographics and comorbidities (in column 2). The simulated results in Panel B allow τ_d to follow the estimated posterior distribution (i.e. without correcting for overtesting).