American Economic Review 2016, 106(12): 3730–3764
http://dx.doi.org/10.1257/aer.20140260

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

By Jason Abaluck, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh*

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. (*JEL* I11, I13, I18)

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 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

*Abaluck: Yale University, 165 Whitney Avenue, New Haven, CT 06520, and NBER (e-mail: jason.abaluck@yale.edu); Agha: Department of Economics, Dartmouth College, 6016 Rockefeller Hall, Hanover, NH 03755, and NBER (e-mail: Leila.Agha@dartmouth.edu); Kabrhel: Department of Emergency Medicine, Massachusetts General Hospital, Zero Emerson Place, Suite 3B, Boston, MA 02114, and Harvard University (e-mail: ckabrhel@partners.org); Raja: Massachusetts General Hospital and Harvard University (e-mail: asraja@partners.org); Venkatesh: Yale-New Haven Hospital, and Yale University (e-mail: arjun.venkatesh@yale.edu). 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 University, Stanford University, University of Houston, and Yale University. Funding for this work was provided by NIA grant T32-AG0000186 to the NBER. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

[†]Go to http://dx.doi.org/10.1257/aer.20140260 to visit the article page for additional materials and author disclosure statement(s).

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 one-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 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 computerized tomography (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

[1] In other ongoing work, we extend the framework developed here to the case of two-sided selection also studied by Chandra and Staiger (2011). Specifically, we analyze 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.

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 are 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. 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 among 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, Gentzkow, and Williams (2014) find that roughly one-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 physicians' 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 percent 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.

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, 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 percent 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 percent and the number of chest CT scans would fall by 50 percent. 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 percent, primarily by leading to additional testing and appropriate diagnosis of patients with a PE.

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

I. Background on Pulmonary Embolism 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

[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.

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 one-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 percent 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 online Appendix 1, 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 unnecessary treatment

[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 (Well et al. 1995, 1998, 2000).

with anticoagulants, incurring financial costs and creating significant risk of bleeding. In addition, there is an estimated 0.02 percent chance of a severe reaction to the contrast, which then carries a 10.5 percent 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 online Appendix 1.

II. Data and Summary Statistics

In this section, we describe our sample construction and present summary statistics that provide initial evidence of variation in physicians' use of tests and misweighting of patient observable characteristics.

A. *Medicare Claims Data*

Using a 20 percent 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.

To identify patients evaluated in the emergency department (ED), we use physician-submitted Medicare Part B claims for ED 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 on the 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

[5] In particular, we identify patients based on current procedural terminology (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 one-half 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.

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 online Appendix 2.

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 ED 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 online Appendix 3.

In addition to measuring whether patients were tested and the testing outcome, we also document a number of characteristics which allow us to predict the patient's propensity to be 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.[8]

Summary statistics are reported in online Appendix Table 1. 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 percent of them are tested with a chest CT scan with contrast. Among tested patients, 6.9 percent of them receive a positive test, i.e., are admitted to the hospital with a diagnosis of PE.

B. *Reduced-Form Evidence of Heterogeneity in Testing Intensity*

Before describing our model, we consider reduced-form evidence of heterogeneity in doctors' testing behavior. Average testing rates vary tremendously across doctors in our sample, from an average of 1.7 percent of ED patients tested in the lowest decile of physician test rates to an average of 8.2 percent of ED patients tested in the highest decile of physician test rates. Does this variation reflect differences in doctor behavior for patients with similar PE risk, or differences in patient PE risk for physicians with similar testing intensities?

[8] PE-specific risk factors include 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.

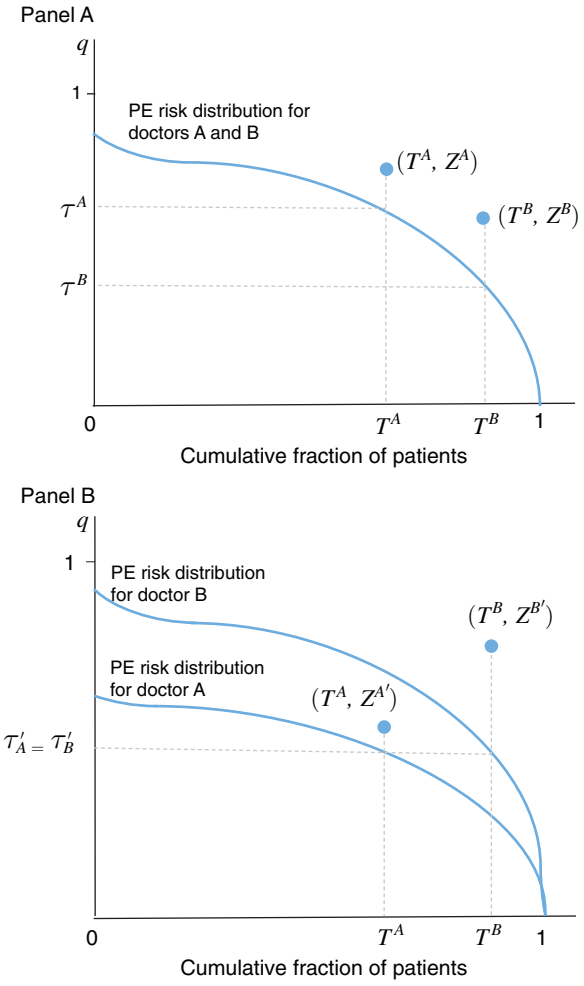

FIGURE 1. STYLIZED RELATIONSHIP BETWEEN TESTING THRESHOLDS, TESTING RATES, AND TEST YIELDS

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.

We can separate these hypotheses by comparing rates of positive tests conditional on testing behavior. For illustration, we have sketched a stylized picture of the testing decisions in Figure 1. Each panel shows the testing decisions and yields for two physicians, with Doctor A testing fewer patients than Doctor B. 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 $(T^A = 2/3, \tau^A = 1/2)$ in panel A, the graph indicates that $2/3$ of patients have a risk of PE that equals or exceeds $1/2$. (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 that 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 threshold 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 scatter plot 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 panel A of Figure 2. 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 are also testing more low-risk patients compared to peer physicians.

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 III below.

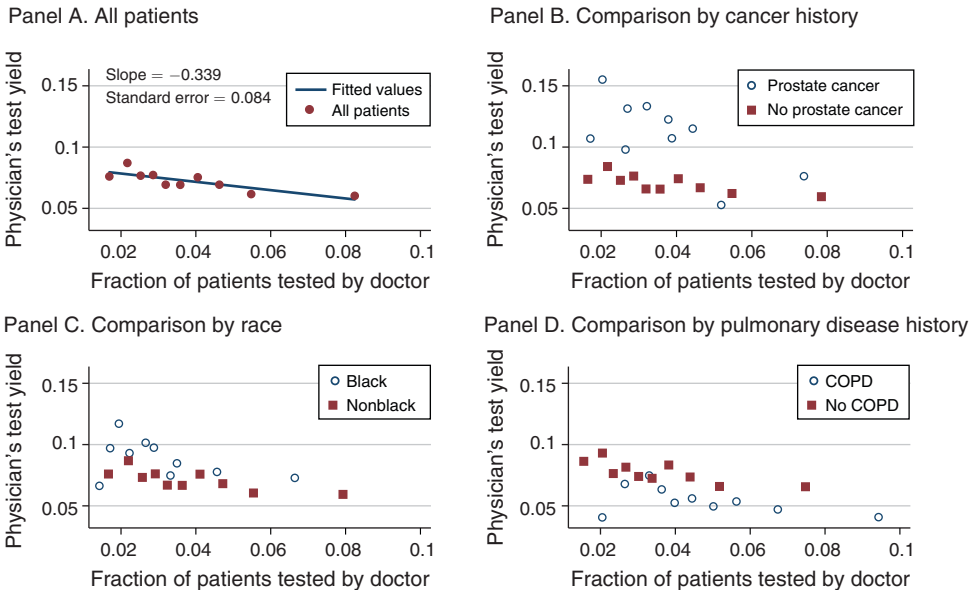

FIGURE 2. BINNED SCATTER PLOT OF PHYSICIAN TEST YIELD BY FRACTION OF PATIENTS TESTED

Notes: Figure displays a binned scatter plot 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.

C. *Reduced-Form Evidence of Misweighting 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.

The frequency of PE testing is slightly lower among patients with a history of prostate cancer (3.7 percent) compared to the rest of the population (3.8 percent). However, it turns out that among tested individuals, prostate cancer patients are over 50 percent more likely to be diagnosed with PE than patients with no such history. In panel B of Figure 2, 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. Paralleling arguments made in the previous section, if patients with a given comorbidity have higher yield they should also be tested at higher rates.

Although cancer is a recognized clinical risk factor for PE included in a popular risk score (Wells et al. 1995, 1998, 2000), patients with a history of prostate

Table 1—Summary Statistics Illustrating Potential Misweighting of Risk Factors

	Fraction tested (1)	Test yield (2)
Selected candidates for underweighting		
Prostate cancer (CCW)	0.0370	0.1019
No prostate cancer (CCW)	0.0380	0.0677
Black	0.0313	0.0851
Nonblack	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 overweighting		
Chronic obstructive pulmonary disease (CCW)	0.0466	0.0524
No chronic obstructive pulmonary disease (CCW)	0.0360	0.0742
Ischemic heart disease	0.0376	0.0566
No ischemic heart disease	0.0382	0.0786
Atrial fibrillation	0.0317	0.0520
No atrial fibrillation	0.0388	0.0713

Notes: Table reports summary statistics for selected comorbidities to motivate the examination of misweighting. Column 1 reports average rates of testing for patients with and without the listed conditions. Column 2 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.

Source: Data are from the Medicare claims, 2000–2009.

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 in their testing decisions.

In Table 1, 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 suggestive evidence that black patients are undertested given their elevated rates of PE. Panel C of Figure 2 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 fibrillation, 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 percent *less* likely to have a PE detected. Panel D of Figure 2 shows the test rates are substantially higher and yields lower for patients 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. 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.

III. 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.

To recover a measure of physician practice style that is purged of variation due to different patient populations, we apply an insight from Chandra and Staiger (2011)—henceforth, CS—which builds upon classical selection models developed by Heckman (1979) and Heckman and Macurdy (1980). In Section IIIA, we lay out the CS model with one small change, 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 IIIB, 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 IIIC, we discuss how physician thresholds, misweighting, and the degree of selection on unobservables can be jointly identified. In Section IIID, we provide further details on how our model is estimated.

A. *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:

$$q_{id} = x_{id}\beta + \alpha_d + \eta_{id}, \tag{1}$$

where x_{id} are observed patient characteristics (which we assume throughout are normalized to have mean zero 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 α_d allows the population risk of PE to vary across doctors in ways that are not captured by the included patient covariates.[9]

[9] CS interpret α_d to reflect variation in doctor 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

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 IVD.) 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

$$Test_{id} = 1 \leftrightarrow q_{id} = x_{id}\beta + \alpha_d + \eta_{id} > \tau_d, \tag{2}$$

which implies that

$$\Pr(Test_{id} = 1) = f(x_{id}\beta + \alpha_d - \tau_d), \tag{3}$$

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 $I_{id} = x_{id}\beta + \alpha_d - \tau_d$.

The threshold τ_d is our measure of physician testing 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 ordinary least squares (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 test outcomes as follows:

$$\begin{aligned} E(q_{id}|Test_{id} = 1) &= E(Z_{id}|q_{id} > \tau_d) = x_{id}\beta + \alpha_d + E(\eta_{id}|q_{id} > \tau_d) \\ &= x_{id}\beta + \alpha_d + h(x_{id}\beta + \alpha_d - \tau_d) \\ &= \tau_d + \lambda(I_{id}), \end{aligned} \tag{4}$$

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.

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 percent, then for marginal patients (with the minimum observed value of I_{id} among tested patients), the positive test probability is exactly 3 percent. The probability 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 scatter plot of testing rates and test yields described in Section IIB can provide some intuition for understanding this model. If all variation across doctors in testing behavior were driven by patient PE risk, then physicians with higher average testing propensities would 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.[10] 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 IIB, 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 IIIC.

B. *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 underweight or overweight 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

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

[10] 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.

while the actual probability remains

$$q_{id} = x_{id}\beta + \alpha_d + \eta_{id}. \tag{6}$$

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

$$x_{id}\beta' + \alpha_d + \eta_{id} > \tau_d + \alpha_d - \alpha'_d. \tag{7}$$

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.[11]

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:

$$\begin{aligned} E(Z_{id}|Test_{id} = 1) &= E(q_{id}|\, q'_{id} > \tau_d) \\ &= E(q'_{id}|\, q'_{id} > \tau_d) + x_{id}(\beta - \beta') \\ &= \tau_d + x_{id}(\beta - \beta') + \lambda(I'_{id}). \end{aligned} \tag{8}$$

The derivation above 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.

C. *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 misweighting.

[11] Note that an analogous argument implies that it is without loss of generality to allow testing thresholds to vary with observables. That is, suppose $\tau_d = \bar{\tau}_d + x_{id}\gamma$. 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.

The CS model is essentially the model we outline in Section IIIA—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} 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 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.[12] We exploit the fact that τ_d can be directly estimated for physicians testing patients we can identify as marginal.[13] 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:

$$QQ_d = \tau_d + E_{m,d}(x_{id} | Test_{id} = 1)(\beta - \beta'). \tag{9}$$

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 physicians' 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 online Appendix 4. 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 identify τ_d for other doctors.

[12] 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.

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

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 relationship between α_d and test outcomes will flexibly identify the shape of the $\lambda(\cdot)$ function.

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 online Appendix 7, we consider one such model and show that it does not materially impact our results.

Identification of misweighting follows from the ignorability assumption: 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 among 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 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 IIC, 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.

D. *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:

$$Test_{id} = 1 \leftrightarrow x_{id}\beta' + \theta'_d + \eta_{id} \geq 0. \tag{10}$$

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 IVD, 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, $\eta_{id} \sim U[v-\eta, v+\eta]$. 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 v (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: first, 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. Second, 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 online Appendix 7 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 online Appendix 7.

In online Appendix 4, we show that this distributional assumption implies

$$\Pr(Test_{id} = 1) = \max\left\{0, \frac{p}{2} + \frac{p\,(I'_{id} + v)}{2\eta}\right\}, \tag{11}$$

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

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, $E(\eta_{id} | \eta_{id} > -I'_{id}) = \frac{\eta - I'_{id} + v}{2}$. As shown in more detail in online Appendix 4, this implies that

$$E(Z_{id} | Test_{id} = 1) = \tau_d + x_{id}(\beta - \beta') + \frac{\eta \tilde{I}'_{id}}{p}. \tag{12}$$

As discussed in Section IIIC, 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 online Appendix Table 4 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 (12) yields

$$Y_{id} = (1 - M_d)\,\tau_d + \frac{\eta \tilde{I}'_{id}}{p} + X_{id}(\beta - \beta') + \epsilon_{id}\,, \tag{13}$$

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 (13) 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 nonmarginal 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 (13) for doctors with nonmarginal patients, estimating the model jointly for all doctors uses additional information about the relative value of the parameters for doctors with nonmarginal patients; this increases the precision of the estimates but has little impact on the magnitude of the coefficients.

Least squares estimation of equation (13) will allow us to recover the constant $\frac{\eta}{p}$ and doctor fixed effects τ_d for nonmarginal 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 online Appendix 5.[14] 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 online Appendix 5.

IV. Results

In this section, we report results of the estimation strategy described in Section IIID above. First, we describe the recovered distribution of physician testing thresholds. Then, we report results on which risk factors are underweighted and

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

overweighted 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.

A. *Distribution and Correlates of Physician Testing Thresholds*

After estimating the model laid out in Section III and applying the empirical Bayes adjustment, we find the mean value of τ_d is 0.056 and the standard deviation is 0.054.[15] 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 percent. 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 our sample is only 6.9 percent, it is likely that this variation in testing thresholds may affect testing decisions for many patients.

In online Appendix Table 2, we consider OLS and feasible generalized least squares (FGLS) regressions of τ on doctor, hospital, and regional level variables. We find that less experienced physicians and physicians practicing in higher spending regions have lower test thresholds, but no evidence that tort reform, hospital nonprofit status, or medical school ranking are relevant. 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 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, Rockoff, and Staiger 2014).

B. *Identifying Misweighted Comorbidities*

Next, we explore physicians' misweighting of observable PE risk factors. As outlined in Section IIIB, we focus on measuring aggregate misweighting: factors which appear to be systemically under- or overweighted 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 overweighting 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 2 and online Appendix Table 3. 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

[15] 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.

Table 2—Comorbidities with Significant Misweighting: Impact of Comorbidity on Testing Decisions and Estimated Misassessment of PE Risk

	Marginal effect from testing equation (1)	Misassessment of PE risk (2)	SE of misassessment (3)	*t*-statistic of misassessment (4)
Underweighted risk factors				
Prior hospital visit within 30 days	−0.0094	0.1070	0.0121	8.8430
Prior hospital visit within 7 days	−0.0041	0.1128	0.0130	8.6769
Prostate cancer (CCW)	0.0014	0.0298	0.0048	6.2083
Cancer metastasis (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
Rheumatoid 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
Osteoporosis (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 online Appendix Table 3, 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.

Source: Data are from the Medicare claims, 2000–2009.

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 (12)). 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.[16]

[16] 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

We find evidence of substantial under- and overweighting 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 overestimates. 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.[17] 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.[18]

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 $\tilde{I}'_{id}$.

[17] Prostate cancer, metastatic cancer, endometrial cancer, and colorectal cancer all have significant underweighting.

[18] 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

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 (Goldhaber et al. 1983; Myllnen et al. 1985; Matta et al. 2009). A complete list of underweighted risk factors is reported in the top panel of Table 2.

A number of different conditions that mimic the symptoms of PE appear on the list of overweighted comorbidities: 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.

Turning to demographic variables, we find evidence that black patients are undertested. They are less likely to be tested for PE than nonblack patients, despite the fact that they are at higher risk of PE. 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.

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.[19]

C. *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

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.

[19] 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.

of patients tested as we did in Figure 2, we bin physicians by the structural analogue: the average estimated testing propensity $\hat{\hat{I}}'_{id}$ across their patients. Recall the observation from the reduced-form analysis in Section IIB 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 online Appendix 6.

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.[20]

Finally, we investigate how misweighting impacts the 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 online Appendix 6.

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 VC.) Inframarginal patients are likely to be tested even with misweighting, but the set of marginal patients changes. Some patients who are less

[20] If we graphed testing propensities versus 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 nonmonotonicity 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 among tested patients due to the convexity of the relationship between I_{id} and positive testing rates at the individual level.

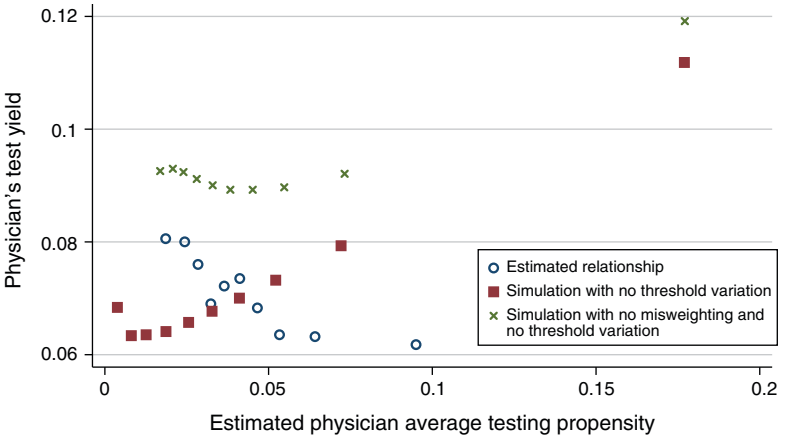

FIGURE 3. BINNED SCATTER PLOT OF PHYSICIAN TEST YIELD BY TESTING PROPENSITY INDEX: ESTIMATION RESULTS AND SIMULATIONS

Notes: Figure displays a binned scatter plot 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 plot the relationship between physicians' actual test yields and physicians' average $\tilde{I}'_{id}$. The 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 display 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.

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.

D. *Robustness*

The results discussed in the previous sections depend on a number of modeling assumptions. Two crucial assumptions underlie 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 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.

Details of these alternative estimation procedures and results are presented in online Appendix 7 and online Appendix Tables 4 and 5. Taken together, these robustness checks 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 is driven by differences in practice styles. Further, doctors are misassessing patient PE risk by similar amounts in percentage point terms across all models.

V. 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.

A. *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: $NU\, q_{id} - c > 0$ 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). 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. In online Appendix 8, we lay out a simple extension to the model that adjusts the benefits of testing to account for both type I and type II errors.

Table 3 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

Table 3—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	0.025VSL	MB	Lessler et al. 2010
Medical cost of treating PE	0.0017VSL	MC	Lessler et al. 2010
Financial cost of testing	$300	c	Estimated from Medicare claims
Financial cost of PE treatment	$2,800	CT	Estimated from Medicare claims

Note: Calibrated parameters of the model applied in welfare simulations reported in Section V.

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.[21] We estimate the financial cost of testing and the financial cost of PE treatment directly from our Medicare claims data. Online Appendix Table 8, 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. Following the estimate from the medical literature, we report results with a false positive rate of 4 percent as our preferred welfare calibration, but also show the welfare implications of assuming a 3 percent or 0 percent false positive rate.[22] 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 4 reports the optimal testing threshold τ^* under these calibration assumptions. With a false positive rate of 4 percent, we find physicians should optimally test all patients with an ex ante likelihood of a positive test greater than or equal to 6.2 percent. The optimal threshold decreases to 5.0 percent at a false positive rate of 3 percent; at the (unlikely) extreme of no false positive test results, the optimal threshold falls to 1.5 percent.

B. *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 percent of doctors overtesting under our preferred calibration assumptions. Second, unlike the overtesting case, we find that the welfare loss due to undertesting 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

[21] 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.

[22] 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 percent (Stein et al. 2006).

TABLE 4—PATIENT WELFARE WITH OBSERVED TESTING THRESHOLDS VERSUS 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 overtesting (percent)	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 (percent)	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	1,318.7	1,045.5	1,368.3	1,752.8	1,770.5
Net benefits per test ($)	188.1	609.6	426.7	693.0	1,134.0	1,149.3

Notes: We compare testing behavior and social welfare under the estimated posterior distribution of physician testing thresholds τ_d (columns 1, 3, and 5) to simulated behavior assuming all physicians with thresholds below the calibrated optimum are reassigned to the optimal testing threshold of $\tau_d = \tau^*$ (columns 2, 4, and 6). 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.

welfare calculations, a specific distributional assumption is required. For some distributions of τ_d, even a small number of doctors undertesting can lead to large welfare losses if the right tail of the τ_d distribution is sufficiently thick.

To calculate the welfare cost of overtesting, 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.[23] Table 4 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. Our earlier results found that the average doctor tests a patient if the probability of a positive test exceeds 5.6 percent. We want to know: how would testing behavior change for each physician if all physicians with testing thresholds below $\tau^* = 6.2$ percent instead adopted a threshold of 6.2 percent? 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 $\frac{\eta}{p}$, and how the probability of a positive test conditional on testing changes for each observation. More details are provided in online Appendix 8.

Results are shown in Table 4, under a series of different assumptions about the false positive rate. At a false positive rate of 4 percent (the estimate in the medical literature), 84 percent 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 percent threshold probability of a positive test the calibration suggests is optimal. At a false positive rate of 3 percent, the proportion of doctors overtesting falls to 67.2 percent. To illustrate the

[23] Note that τ_d is bounded below at the false positive rate.

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 percent and only 10 percent of physicians are overtesting on the margin, i.e., have a testing threshold lower than 1.5 percent. At a false positive rate of 3 percent or 4 percent, eliminating overtesting would decrease the total number of patients tested by more than 30 percent or 50 percent, respectively.

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. As detailed in Table 4, 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 4, 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. However, Finkelstein, Gentzkow, and Williams (2014) find that variation in patient demand (i.e., both patient preferences and medical needs) explains only 14 percent of the regional variation in spending on imaging, suggesting a very limited role for patient preferences in driving 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 online Appendix Table 8 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 percent to 14.3 percent. 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-à-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 percent, 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 percent, 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 percent, eliminating most overtesting but implying large amounts of undertesting. 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.

C. *Welfare Impact of Eliminating Misweighting of Patient Risk Factors*

Table 5 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 undertesting and overtesting. We report results at our preferred calibration of the false positive rate, 4 percent; the welfare consequences of eliminating misweighting would be even larger at lower false positive rates. Technical details of this exercise are described in online Appendix 8.

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 of 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 column 3 of Table 5. 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 percent to 4.3 percent, 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 four times as large as the welfare loss from overtesting ($8.1 million), even in the model with the highest rate of false positives.

Table 5—Patient Welfare with Observed Misweighting versus in Simulations with No Misweighting

	False positive rate of 4 percent		
	Actual testing decisions (1)	No misweighting, simulation without cross-validation (2)	No misweighting, simulation with cross-validation (3)
Description of results			
Percent of patients tested	3.8	4.3	4.3
Number of patients tested	71,314	81,410	79,734
Test yield among tested patients (percent)	7.0	9.2	8.6
Number of positive tests detected	5,019	7,526	6,872
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	1,311.3	1,213.1
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 (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 (column 2). The simulated results in the bottom panel allow τ_d to follow the estimated posterior distribution (i.e., without correcting for overtesting).

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 online Appendix Table 7. 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 percent 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 percent of the total potential gains. These covariates are both substantially misweighted and common enough to induce large welfare consequences.

Panel B of online Appendix Table 8 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.

Undiagnosed PE is thought to be a major public health problem, with the Office of the Surgeon General (2008) estimating that approximately one-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 percent sample of patients enrolled in Medicare Parts A and B over a 10-year period, and the numbers reported in Tables 4 and 5 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 five to account for the entire population of Medicare fee for service enrollees, then adjust to account for the 28 percent of Medicare patients who enroll in a Medicare Advantage plan, and finally divide by ten 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.

VI. 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 overtesting 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 physicians' 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 percent 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 overweighted 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 misallocation 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.

REFERENCES

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh. 2016. "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care: Dataset." *American Economic Review.* http://dx.doi.org/10.1257/aer.20140260.

Chandra, Amitabh, and Douglas O. Staiger. 2010. "Identifying Provider Prejudice in Healthcare." National Bureau of Economic Research Working Paper 16382.

Chandra, Amitabh, and Douglas O. Staiger. 2011. "Expertise, Oversure and Underuse in Healthcare." Unpublished.

Coco, Andrew S., and David T. O'Gurek. 2012. "Increased Emergency Department Computed Tomography Use for Common Chest Symptoms without Clear Patient Benefits." *Journal of the American Board of Family Medicine* 25 (1): 33–41.

Costantino, Mary M., Geneva Randall, Marc Gosselin, Marissa Brandt, Kristopher Spinning, and C. David Vegas. 2008. "CT Angiography in the Evaluation of Acute Pulmonary Embolus." *American Journal of Roentgenology* 191 (2): 471–74.

Currie, Janet, and W. Bentley MacLeod. 2013. "Diagnosis and Unnecessary Procedure Use: Evidence from C-Section." National Bureau of Economic Research Working Paper 18977.

Elixhauser, Anne, Claudia Steiner, D. Robert Harris, and Rosanna M. Coffey. 1998. "Comorbidity Measures for Use with Administrative Data." *Medical Care* 36 (1): 8–27.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2014. "Sources of Geographic Variation in Health Care: Evidence from Patient Migration." National Bureau of Economic Research Working Paper 20789.

Garber, Alan M., and Jonathan Skinner. 2008. "Is American Health Care Uniquely Inefficient?" National Bureau of Economic Research Working Paper 14257.

Goldhaber, Samuel Z., and Henri Bounameaux. 2012. "Pulmonary Embolism and Deep Vein Thrombosis." *Lancet* 379 (9828): 1835–46

Goldhaber, Samuel Z., David D. Savage, Robert J. Garrison, William P. Castelli, William B. Kannel, Patricia M. McNamara, Gherardo Gherardi, and Manning Feinleib. 1983. "Risk Factors for Pulmonary Embolism: The Framingham Study." *American Journal of Medicine* 74 (6): 1023–28.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–61.

Heckman, James J., and Thomas E. Macurdy. 1980. "A Life Cycle Model of Female Labour Supply." *Review of Economic Studies* 47 (1): 47–74.

Iglehart, John K. 2009. "Health Insurers and Medical-Imaging Policy—A Work in Progress." *New England Journal of Medicine* 360 (10): 1030–37.

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics* 6 (1): 801–25.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." National Bureau of Economic Research Working Paper 14607.

Lessler, Adam L., Joshua A. Isserman, Rajan Agarwal, Harold I. Palevsky, and Jesse M. Pines. 2010. "Testing Low-Risk Patients for Suspected Pulmonary Embolism: A Decision Analysis." *Annals of Emergency Medicine* 55 (4): 316–26.

Mamlouk, Mark D., Eric vanSonnenberg, Rishi Gosalia, David Drachman, Daniel Gridley, Jesus G. Zamora, Giovanna Casola, and Sanford Ornstein. 2010. "Pulmonary Embolism at CT Angiography: Implications for Appropriateness, Cost, and Radiation Exposure in 2003 Patients." *Radiology* 256 (2): 625–32.

Matta, Fadi, Ravinder Singala, Abdo Y. Yaekoub, Reiad Najjar, and Paul D. Stein. 2009. "Risk of Venous Thromboembolism with Rheumatoid Arthritis." *Thrombosis and Haemostasis* 101 (1): 134–38.

Molitor, David. 2012. "The Evolution of Physician Practice Styles Evidence from Cardiologist Migration." National Bureau of Economic Research Working Paper 22478.

Mulligan, Casey B., and Yona Rubinstein. 2008. "Selection, Investment, and Women's Relative Wages over Time." *Quarterly Journal of Economics* 123 (3): 1061–110.

Murphy, Kevin M., and Robert H. Topel. 2006. "The Value of Health and Longevity." *Journal of Political Economy* 114 (5): 871–904.

Myllynen P., M. Kammonen, P. Rokkanen, O. Bostman, M. Lalla, and E. Laasonen. 1985. "Deep Venous Thrombosis and Pulmonary Embolism in Patients with Acute Spinal Cord Injury: A Comparison with Nonparalyzed Patients Immobilized Due to Spinal Fractures." *Journal of Trauma* 25 (6): 541–43.

Nelson, Alan. 2002. "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care." *Journal of the National Medical Association* 94 (8): 666–68.

Office of the Surgeon General and National Heart, Lung, and Blood Institute. 2008. "The Surgeon General's Call to Action to Prevent Deep Vein Thrombosis and Pulmonary Embolism." Rockville, MD: Office of the Surgeon General.

Rahimtoola, Ali, and James D. Bergin. 2005. "Acute Pulmonary Embolism: An Update on Diagnosis and Management." *Current Problems in Cardiology* 30 (2): 61–114.

Rao, Vijay M., and David C. Levin. 2012. "The Overuse of Diagnostic Imaging and the Choosing Wisely Initiative." *Annals of Internal Medicine* 157 (8): 574–76.

Stein, Paul D., Sarah E. Fowler, Lawrence R. Goodman, Alexander Gottschalk, Charles A. Hales, Russell D. Hull, Kenneth V. Leeper Jr. et al. 2006. "Multidetector Computed Tomography for Acute Pulmonary Embolism." *New England Journal of Medicine* 354 (22): 2317–27.

Venkatesh, Arjun K., Jeffrey A. Kline, D. Mark Courtney, Carlos A. Camargo Jr., Michael C. Plewa, Kristen E. Nordenholz, Christopher L. Moore et al. 2012. "Evaluation of Pulmonary Embolism in the Emergency Department and Consistency with a National Quality Measure: Quantifying the Opportunity for Improvement." *Archives of Internal Medicine* 172 (13): 1028–32.

Wells, Philip S., David R. Anderson, Marc Rodger, Jeffrey S. Ginsberg, Clive Kearon, Michael Gent, A. G. Turpie et al. 2000. "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* 83 (3): 416–20.

Wells, Philip S., Jeffrey S. Ginsburg, David R. Anderson, Clive Kearon, Michael Gent, Alexander G. Turpie, Janis Bormanis et al. 1998. "Use of a Clinical Model for Safe Management of Patients with Suspected Pulmonary Embolism." *Annals of Internal Medicine* 129 (12): 997–1005.

Wells, Philip S., Jack Hirsh, David R. Anderson, Anthony W. Lensing, Gary Foster, Clive Kearon, Jeffrey Weitz et al. 1995. "Accuracy of Clinical Assessment of Deep-Vein Thrombosis." *Lancet* 345 (8961): 1326–30.

Wennberg, John E., Megan McAndrew Cooper, Thomas A. Bubolz, Elliott S. Fisher, Alan M. Gittelsohn, David C. Goodman, Jack E. Mohr et al. 1996. *The Dartmouth Atlas of Health Care in the United States.* Chicago: American Hospital Publishing, Inc.