NBER WORKING PAPER SERIES

NEGATIVE TESTS AND THE EFFICIENCY OF MEDICAL CARE: WHAT DETERMINES HETEROGENEITY IN IMAGING BEHAVIOR?

Jason Abaluck
Leila Agha
Christopher Kabrhel
Ali Raja
Arjun Venkatesh

Working Paper 19956
http://www.nber.org/papers/w19956

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2014

Thanks to Brian Abaluck, Joe Altonji, Joshua Aronson, 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, Blair Parry, Michael Powell, Constana Esteves-Sorenson, Ashley Swanson, Bob Town, and Heidi Williams as well as seminar participants at AHEC 2012, AEA meeting 2013, Boston University, Cornell, HEC Montreal, IHEA 2013, the National Bureau of Economic Research, NIA Dartmouth research meeting, the National Tax Association annual meeting, Northwestern, Stanford, University of Houston, and Yale. Funding for this work was provided by NIA Grant Number T32-AG0000186 to the NBER as well as internal research funds at Yale University and Boston University. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2014 by Jason Abaluck, Leila Agha, Christopher Kabrhel, Ali Raja, and Arjun Venkatesh. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Negative Tests and the Efficiency of Medical Care: What Determines Heterogeneity in Imaging Behavior?

Jason Abaluck, Leila Agha, Christopher Kabrhel, Ali Raja, and Arjun Venkatesh

NBER Working Paper No. 19956
March 2014
JEL No. I0,I12

ABSTRACT

We develop a model of the efficiency of medical testing based on the frequency of negative CT scans for pulmonary embolism. The model is estimated using a 20% sample of Medicare claims from 2000-2009. We document enormous heterogeneity in testing conditional on patient population. Less experienced physicians and those practicing in high spending areas test more low-risk patients. Assessing the efficiency of current practices requires calibration assumptions regarding the costs of testing, the benefits of treatment and the likelihood of false positives. While we cannot tell whether any particular testing decision was mistaken in the context of our model, we find that collectively–given these additional calibration assumptions–there are systematic differences between doctor testing practices and the recommendations of our model of optimal testing. According to our model, 90-99% of doctors test even when costs exceed expected benefits; optimal testing thresholds would increase social welfare by 20-35%. Shifting doctor practice to weight risk factors differently could increase net welfare in our model by 275%.

Jason Abaluck
Yale School of Management
Box 208200
New Haven, CT 06520-8200
and NBER
jason.abaluck@yale.edu

Leila Agha
School of Management
Boston University
595 Commonwealth Avenue
Boston, MA 02215
and NBER
lagha@bu.edu

Christopher Kabrhel
Department of Emergency Medicine
Massachusetts General Hospital
55 Fruit St, Clinics Building 115
Boston, MA 02114
ckabrhel@partners.org

Ali Raja
Department of Emergency Medicine
Brigham and Women's Hospital
75 Francis St
Boston, MA 02115
asraja@partners.org

Arjun Venkatesh
Yale School of Medicine
Department of Internal Medicine
arjun.venkatesh@yale.edu

1 Introduction

Many have argued that current medical practice involves large amounts of wasteful spending, with little cross-sectional correlation between regional health spending and quality of care (Wennberg, Cooper, et al. 1996). But determining the best approach to lower costs and maintain quality depends critically on the nature of the inefficiency: is the problem one of allocative inefficiency, i.e. spending to the "flat of the curve," or productive inefficiency, i.e. misuse of medical resources (Garber and Skinner 2008)? And more fundamentally, does heterogeneity in medical spending imply inefficiency of any sort or is this heterogeneity fully explicable by unobserved differences in patient populations?

In this paper, we develop an econometric framework for evaluating whether the wide variation in the use of CT imaging across doctors is due to heterogeneity in patients' benefits from testing or heterogeneity in physicians' practice styles (i.e. differences in behavior when treating identical patients). We assume that doctors will order a CT scan to test for pulmonary embolism (PE) if the expected benefits to the patient minus the cost of the test exceed a doctor-specific testing threshold. This threshold is our patient invariant measure of physician practice style 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 random assignment of patients to physicians or estimates of potentially heterogeneous returns to medical treatment for each patient. Prior research, including Chandra and Staiger (2011) and Currie and MacLeod (2013), has argued that reliable estimates of treatment effects can be obtained using detailed chart data to control for all patient characteristics observable to doctors, but such data is typically only available in limited samples. This stumbling block makes it difficult to investigate both the extent and the determinants of healthcare overuse or underuse.

A key insight of this paper is that the *ex post* value of a medical 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. Physicians with identical testing thresholds may have different rates of negative tests if their patient population varies either in the *ex ante* risk of pulmonary embolism or in the benefits of treatment conditional on a positive test. Our model accounts for these forms of patient heterogeneity and shows how to recover physicians' testing thresholds. Using these estimated testing thresholds, we investigate the role of medical training, malpractice environment, hospital characteristics and regional factors in shaping practice styles.

The model also allows identification of whether doctors are misweighting observable patient risk factors in selecting which patients to test for PE. By comparing how observables predict physicians' testing decisions to how those same observables predict rates of positive tests amongst tested patients, we can identify whether physicians are mistakenly assessing the risk of PE associated with patient demographics and comorbid conditions.

Given additional calibration assumptions about the cost of the test, the benefits of treating PE and the likelihood of false positives, the testing threshold tells us the degree of allocative inefficiency: whether doctors are overtesting or undertesting from a social standpoint. For example, a negative testing threshold implies that doctors test even if the costs of the test exceed the medical benefits to the patient. The calibration also allow us to assess the degree of productive inefficiency from

physician misweighting of patient risk factors. Can greater benefits be realized at the same cost by more carefully choosing which patients to test?[1]

Our model builds on the framework of Chandra and Staiger (2011) who use a structural model to estimate the medical returns to heart attack treatment and separately identify heterogeneity in physician behavior from differences in patient population. Their framework relies on detailed patient-specific estimated treatment effects. We adapt their model to the case of medical testing, which allows us to investigate overuse in a national sample of claims data, and extend it to allow for additional heterogeneity in the distribution of patient risk and to model physicians' diagnostic errors.

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; however, their framework does not identify overuse or underuse of medical care nor do they evaluate whether doctors place the right relative weights on observables. We consider a close relative of the Currie and Macleod model in Section 6.2 where we consider the role of diagnostic skill in the medical testing context. Molitor (2012) finds that environmental factors explain much of the variation in physician's rates of cardiac catheterization; we build upon this insight by analyzing which environmental and physician-specific variables predict variation in physicians' practice styles.

We apply our model to analyze CT scans that test for pulmonary embolism (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 most common tool for diagnosis of PE. Despite the mortality risk from untreated PE, PE CT scans are commonly thought to be overused in the emergency care setting given their substantial financial costs and medical risks. The American College of Radiology targeted PE CT as a key part of the *Choosing Wisely* campaign aimed to reduce overuse of medical services (American College of Radiology 2012).

We analyze 2.5 million emergency department visits drawn from a 20% sample of Medicare claims data, 2000-2009. We observe whether each patient is tested with a chest CT, and whether this test leads to diagnosis of pulmonary embolism. After accounting for differences in patients' risk for PE, we uncover enormous heterogeneity in doctors' testing thresholds. Less experienced doctors and doctors in higher spending regions tend to have lower thresholds at which they deem CT imaging worthwhile.

Applying calibration assumptions, we find in the context of our model that 90-99% of doctors are overtesting in the sense that for their marginal patients, the costs of testing exceed the benefits. We also use the model to conduct several welfare analyses. If all doctors tested only when the costs exceeded the benefits in our model, the total benefits to patients from chest CTs would increase by 30% and the number of chest CT scans would fall by 40% in our preferred calibration. Our model

[1]We are defining allocative and productive inefficiency from the standpoint of production functions with spending on CT scans as the input and patient health status as the output. With "overtesting", one is testing to the flat of the curve, where the marginal health returns to additional spending are very low; this is an allocative inefficiency. With misweighting, a higher production function—producing greater health gains for a given level of spending—is achievable if doctors would correctly weight observable risk factors; misweighting thus creates productive inefficiency.

also suggests that doctors underweight the importance of certain risk factors, including recent prior hospitalizations, obesity and a history of pulmonary disease, in deciding which patients to test. Weighting these observables in the way our model suggests is optimal would increase patient welfare by 275%, primarily by leading to additional testing and appropriate diagnosis of affected patients.

The paper is organized as follows. Section 2 provides some background on chest CT scans and especially chest CT scans for pulmonary embolism, the test which is the focus of our analysis. Section 3 describes the data available to us and the assumptions needed to identify positive and negative tests. Section 4 lays out our structural model of testing behavior and derives an equation relating the indicator for positive and negative tests to doctors' testing thresholds. Section 5 describes how we estimate the model and reports some results. Section 6 conducts the various welfare analyses described and section 7 concludes.

2 Background on PE CTs and Data

We study testing behavior in the context of chest CT scans performed in the emergency department (ED) to detect pulmonary embolism. A pulmonary embolism 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 pulmonary embolism depends on the severity and has been estimated to be 2.5% within three months for a small PE (Lessler et al. 2010), with most of the risk concentrated within the first hours after onset of symptoms (Rahimtoola and Bergin 2005). Accurate diagnosis of PE is necessary for appropriate follow-up treatment; even high risk patients are unlikely to be treated presumptively.

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

2.1 CT indications and guidelines

The symptoms of pulmonary embolism 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 factors before determining whether to pursue a workup for PE, including the following: 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. Because PE is an acute event with a sudden onset, the workup must be completed urgently and knowing the results of previous CT scans is not a critical part of the evaluation of PE. Despite these guidelines, 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 setting.

A CT angiogram is the standard diagnostic tool for pulmonary embolism. The average allowed charge in the Medicare data is around $320 per PE CT when the bill is not covered by a capitation payment. However, the emergency medicine physician with the responsibility of deciding whether to order a CT scan typically receives no direct financial remuneration from the scan's performance. Payment goes to the radiologist for interpreting the scan and to the hospital for the technician and capital investment required to perform the scan. The emergency department doctor has, at best, 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.

There may be additional opportunity costs of scanning patient, 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 over-testing in our data.

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

2.2 Physician decision tree & value of a negative CT scan

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 pulmonary embolism—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.

The flowchart depicted in Figure 1 shows a typical clinical pathway for a patient who may receive a chest CT. The most common symptom that leads to the consideration of PE as a diagnosis is chest pain. This is a nonspecific symptom that could also indicate a cardiac problem, pneumonia, or a number of other conditions. Blood oxygen tests and an EKG are likely to be performed immediately at the bedside, and if they suggest a cardiac problem, the patient will receive a more complete cardiac workup.

If cardiac conditions are ruled out, the doctor may then be considering pneumonia, pleural effusion, and pulmonary embolism as possible diagnoses. A chest x-ray and D-dimer blood test would be the typical next steps. A chest x-ray is a low cost test with low levels of radiation exposure and little medical risk; it is highly effective at diagnosing pneumonia and pleural effusion, which are more common than PE. If the x-ray is negative, then the physician may become more concerned

about risk of PE, since other more common conditions causing chest pain have been ruled out. A chest x-ray is a commonplace and recommended antecedent to a CT scan; the popular Geneva risk scoring system for evaluating whether patient's PE risk necessitates a CT scan includes chest X-ray findings among the seven risk factors used to calculate the score.

At this point, the physician may consider ordering a D-dimer, an inexpensive blood test that provides further information about a patient's risk of PE. A low-risk result on the D-dimer suggests the patient does not have a PE and the physician may forego a CT scan. A positive D-dimer result is not diagnostic of PE, but suggests an elevated probability of this condition. At this point, the physician would consider ordering a CT scan. Over our study period, the popularity of the D-dimer as an additional screening tool for PE was on the rise. Although we cannot observe the use of the D-dimer in our data, variation in D-dimer utilization is one mechanism by which physician CT ordering behavior may vary.

The physician will typically order a chest CT after ruling out these common causes of chest pain. A chest CT with contrast is useful for diagnosing pulmonary embolism, but otherwise adds little new information that may be helpful in diagnosing other possible acute conditions.[2] A positive test will typically lead to a hospital admission and treatment with blood thinners. Imaging is required for diagnosing PE; even high risk patients have a relatively low probability of PE and PE treatment is medically risky, so it is not a condition that would be treated presumptively without imaging.

A negative CT scan will leave the physician with a broad field of possible alternative diagnoses, including a more subtle cardiac condition, sleep apnea, infection, or a false alarm, and the CT scan result will not be helpful in distinguishing between these possibilities. Ruling out a chest CT has only a modest impact on the posterior probabilities of the other conditions that may be causing a patient's symptoms, since the *ex ante* probability of PE is relatively low—even for higher risk patients. For these reasons, the informational value of a negative test is low.

3 Data

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

3.1 Medicare claims data

We begin by identifying all patients evaluated in the emergency department (ED), using physician-submitted Medicare Part B claims for evaluation and management. [3] The physician submitting this claim for evaluation and management is responsible for the patient's emergency care; it would be his

[2] In Appendix C, we provide a detailed discussion of other conditions that can be diagnosed by chest CT and how we empirically address these possibilities.

[3] In particular, we identify patients based on CPT codes for emergency department evaluation and management: 99281, 99282, 99283, 99284, 99285, and place of service 23 (i.e. hospital emergency department).

decision whether or not to order testing for pulmonary embolis. Using physician identifiers, we track the behavior of all doctors who routinely evaluate Medicare patients in the emergency department.

We identify which ED patients are tested for a PE using bills submitted by radiologists for the interpretation of chest CTs with contrast, when the CT is performed within 1 day of the ED visit.[4] Note that 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 indications, including pleural effusion, chest and lung cancers, and traumas. For this reason, we exclude patients from the sample who are coded with a diagnosis related to pleural effusion, chest or lung cancer, and trauma from the sample. Because a chest x-ray is typically the more appropriate diagnostic tool for pneumonia (rather than chest CT scan), and it is not uncommon to screen pneumonia patients for pulmonary embolism, we do not exclude pneumonia diagnoses from our baseline sample. These sample restrictions and alternative conditions are discussed in more detail in Appendix C.

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

In addition to measuring whether patients were tested and the testing outcome, we also document a number of characteristics that allow us to predict the patient's propensity to be diagnosed with a PE, including age, race, sex, and medical comorbodities. We augment the standard set of 30 medical comorbidities (following Elixhauser et al. 1998) to include several measures that are specific to PE risk.[5] These include whether the patient was admitted to the hospital within the past year with a diagnosis of pulmonary embolism, thoracic aortic dissection, abdominal aortic dissection, deep vein thrombosis, and any cause admission to the hospital within 7 days or 30 days. Comorbidities are defined using a one year history of inpatient Medicare claims.

We restrict our sample to hospitals that have billed Medicare for at least 50 CT scans between 2000-2009 in order to exclude institutions without reliable access to a scanner. We also exclude physicians who order three or fewer CT scans over our period, since these doctors provide too little information to accurately estimate physicians' testing thresholds.

3.2 Validating our approach to claims data

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. In particular, we may undercount

[4] We begin by identifying all bills for chest CTs on the basis of CPT codes 71260, 71270, and 71275.

[5] Conditions are defined using a 1-year inpatient medical history, based on Medicare Part A institutional claims. These diagnoses include: coronary heart failure, valvular disease, pulmonary circulation disorder, peripheral vascular disorder, hypertension, paralysis, other neurological disorders, chronic pulmonary disease, diabetes without chronic complications, diabetes with chronic complications, hypothyroidism, renal failure, liver disease, chronic peptic ulcer, HIV and AIDS, lymphoma, metastatic cancer, solid tumor without metastasis, rheumatoid arthritis, coagulation deficiency, obesity, weight loss, fluid and electrolyte disorder, blood loss anemia, deficiency anemias, alcohol abuse, drug abuse, psychoses, depression.

positive tests in the Medicare claims data for two reasons: if patients with PE are not admitted to the hospital; or if patients with PE are admitted but their inpatient bill does not include a diagnosis of pulmonary embolism.

At the two academic medical centers, we found that 90% of patients who test positive for PE in the emergency department were admitted within 1 day. Patients with very small PEs may occasionally be discharged after brief observation and treated with blood thinning agents as outpatients if the PE appeared small on the scan and the patient has no other complicating health conditions; this likely accounts for most of the cases where a test is coded as positive on the basis of patient chart data but no inpatient admission is recorded. Note that this suggests that we are undercounting positive tests precisely for the patient group for whom the benefits of treatment are the lowest.

Amongst patients with positive PE CT scans recorded in chart data who are subsequently admitted to the hospital, 87% have a diagnosis of pulmonary embolism recorded on the bill for their inpatient hospital stay. PE may not be recorded on the bill for two main reasons: the patient may have other medical conditions that are treated during the hospital stay and are reimbursed at a higher rate, such that there is no billing incentive to include PE amongst the inpatient diagnoses; or, the bill may simply be incorrectly coded. In total, 21% of patients diagnosed with PE in the emergency department (ED) do not have an inpatient claim with a PE diagnosis.

Of patients with a negative PE CT scan recorded in their emergency department chart, 1.5% have a diagnosis of pulmonary embolism recorded on the bill for an ensuing hospital stay. In the claims data, we would mistakenly attribute this diagnosis to the ED workup. This error could occur if the patient develops a PE later in his hospital course and receives a subsequent positive CT test, a plausible mechanism given that the immobilization frequently associated with hospital stays is a risk factor for PEs; alternatively, these PE diagnoses could indicate billing errors.

Taken together, these data suggest that of the 6% of CT tests that we code as positive in the Medicare data, 20% of the patients had negative findings on their initial ED PE CT. Of the 94% of tests we code as negative, 1.1% of the patients had positive ED PE CTs. The overall rate of positive tests is almost exactly equal to what it would be if no such coding mistakes were made, since these two types of coding errors offset each other. This suggests that the limitations of this coding algorithm should not contribute to overstatements of the degree of over-testing in our Medicare sample.

3.3 Physician, hospital, and regional data

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

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

specialists.

Hospital characteristics are drawn from the American Hospital Association Annual Survey. We use these data to observe whether the physician typically practices at a for profit hospital or an academic hospital, defined as a hospital with a board certified residency program. Lastly, we identify the hospital referral region (HRRs) in which each patient is treated. HRRs are regional health care markets defined by the Dartmouth atlas to reflect areas within which patients commonly travel to receive tertiary care. There are 306 HRRs in total. Using data from the Dartmouth Atlas, we link each HRR to measures of the overall intensity of treatment of Medicare patients, including spending per beneficiary and measures of end of life care.

3.4 Summary Statistics

There are 2.5 million emergency department visit evaluations in our dataset, after making sample exclusions. Of these patients evaluated in the ED, 3.5% of them are tested with a chest CT scan with contrast. Amongst tested patients, 7% of them receive a positive test, i.e. are admitted to the hospital within 24 hours with a diagnosis of pulmonary embolism.

Summary statistics are reported in Table 1, with results reported separately for patients who do not receive a CT scan (column A), patients who receive a negative test (column B), and patients with a positive test (column C). We observe the testing behavior of over 10,000 physicians, with an average of over 200 ED patients per physician.

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

Patients with negative tests are evaluated by doctors with seven months less experience on average than patients with positive tests. They are also more likely to have been treated in a slightly higher spending region, with regional average per beneficiary spending 1% higher amongst negative tested patients compared to postive tested patients. 40% of patients are evaluated by a doctor who sees a plurality of his patients at an academic medical center, and 31% of patients are evaluated by a physician who attended a top 50 ranked research medical school; these fractions do not vary much across patient groups.

3.5 Reduced Form Evidence of Doctor Preference Heterogeneity

Before describing our model, we consider reduced form evidence of heterogeneity in doctors' testing behavior. In particular, we analyze the following regression:

$$Z_{id} = t_d + X_{id} + \mu_{id} \tag{1}$$

where Z_{id} is in indicator for whether patient i tested by doctor d received a positive CT scan; t_d is a vector of doctor fixed effects; X_{id} are patient demographics and comorbidities; and μ_{id} is an error term. The regression is run on the sample of patients who receive a chest CT.

Higher fixed effects t_d correspond to a higher rate of positive tests amongst tested patients for physician d conditional on the observable risk factors in his patient population. There is considerable spread in the residual probability of a positive test; the standard deviation of this distribution is 10 percentage points.

Some of the variation in the fixed effects may be driven by the fact that we observe a limited number of patients per doctor, so even if all doctors treated the same population of patients and shared the same testing preferences, we may estimate differences in their testing behavior. Using the empirical Bayes correction described in Appendix A to adjust for this sampling variation, we estimate the true standard deviation of physician fixed effects to be 0.04. Given that 7% of patients test positive, this is a substantial amount of variation across doctors in the fraction of positive tests conditional on observables - a doctor one standard deviation above the mean would find 5 times as many positive tests as a doctor one standard deviation below the mean.

This variation in rates of positive testing could be explained by a few factors. First, physicians may vary in their practice styles, with some physicians choosing to test lower-risk patients than others. Second, the patient population each physician treats may vary in its unobservable risk for pulmonary embolism, generating differences in positive testing rates independent of differences in physician practice style. Our model presented below and the estimation approach in the remainder of the paper work to disentangle these mechanisms in order to isolate variation in practice style.

4 Model of Testing Behavior

We will now develop a model of physician's testing decisions and test outcomes that allows us to identify heterogeneity in testing behavior conditional on patient population even when there is systematic unobservable variation in patient PE risk across doctors. With additional assumptions about the costs of testing and the benefits of treatment conditional on a positive test, we can determine which doctors are under- or overusing medical tests. In our framework, a doctor must decide whether or not to test each patient he evaluates in the emergency department with a chest CT, and the econometrician observes both whether each patient is tested and the outcome of each performed test (positive or negative). This framework is adapted from Chandra and Staiger (2011) (hereafter, CS).

In particular, the model separates out three reasons why one physician may test a higher fraction of his patients conditional on patient observable risk factors.

1. *Physician practice style*: A physician may test more frequently because he has a lower net benefit threshold for identical patients; e.g. because of financial incentives which impact the physician's objective function but not social welfare.[6]

[6]One could also interpret τ_d - the physician's net benefit threshold - as the "value of knowing" that one does not have a pulmonary embolism. In contrast to the case of Huntington's disease (Oster, Shoulson, and Dorsey 2011), we find it normatively unpersuasive to include this term as part of welfare because in most cases a pulmonary embolism is ex ante very unlikely and because the rate of false negatives is sufficiently high that even after testing one has only somewhat reduced that probability. While in isolation patients might express concern about any given condition once it is made salient, they cannot consistently have a large willingness to pay to reduce slightly the probability of all unlikely conditions.

2. *Patient PE risk*: A physician may test more frequently because his patients have a higher risk of PE, even after conditioning on observable patient risk factors.

3. *Patient treatment benefit*: A physician may test more frequently because his patients have a greater benefit from treatment, should they turn out to be diagnosed with PE.

We develop a model of physician testing behavior that accounts for each of these three channels, allowing us to distinguish a physician's practice style (reason 1) from the risk and benefit features of the population he treats.

The starting point for our model is the assumption that a doctor tests a patient if the perceived net benefits of testing given all of the information available to him at the time exceed a doctor-specific threshold value. Let B_{id} denote the net benefits if doctor d tests patient i and let τ_d denote this threshold. Then we assume that doctors test if and only if $B_{id} \geq \tau_d$. If τ_d equals 0, then doctors are behaving efficiently because they test only when the net benefits exceed 0. If $\tau_d > 0$, doctors are undertesting, i.e. there are some patients with positive net benefits who they decide not to test; if $\tau_d < 0$ then doctors are overtesting, i.e. there are some patients with negative net benefits whom they test anyway.

The goal of the model will be to recover the threshold values τ_d based on the observed testing decisions (whether or not an evaluated patient is given a chest CT) and the observed rate of negative tests. We will show as in CS that the threshold variables τ_d can be recovered from a regression of the net benefits of testing on doctor fixed effects conditioning on a flexible function of the propensity to test. A key advantage of investigating the efficient use of medical testing as opposed to medical treatment (as in CS) is that the doctor threshold parameters, τ_d, can be recovered without separately estimating the net benefits of treatment for each patient. In order to recover the absolute magnitude of τ_d, we will need to make assumptions about the magnitude of these net benefits although they can in principle be allowed to vary flexibly with observables such as age and patient medical history.

4.1 Intuition

To motivate the full model, consider first a simplified case where we model only reason 1) for potential heterogeneity in doctors' testing behavior: doctors have varying testing thresholds. In particular, assume for the moment that the net benefits of testing are equal to the probability of a positive test, q_{id}, and there is no heterogeneity in the benefits of treatment across patients who tested positive.

Further assume that conditional on observable characteristics, all physicians evaluate patients with the same distribution of PE risk. The probability of a positive test is given by $q_{id} = x_{id}\beta + \eta_{id}$, x_{id} are observables and η_{id} are factors observable to the doctor but not to the econometrician and are distributed i.i.d. across doctors and patients. For example, η_{id} might include symptoms reported by the patient such as chest pain. Lastly, assume that the costs of testing are a known constant c. Doctors will test if $q_{id} - c \geq \tau_d$.

Under these simplifying assumptions, if doctor A tests a greater fraction of patients than doctor B, conditional on observable patient characteristics, then we can immediately infer that doctor A has a lower testing threshold. Specifically, we could estimate the equation $P(test) = f(x_{id}\beta - \tau_d - c)$

and immediately recover τ_d (up to a normalization), without needing to observe testing outcomes (i.e. whether the patients tested positive for PE).

In this simplified model, the only reason why different physicians have different rates of testing conditional on observable patient characteristics is that they have different testing thresholds—there are no *unobservable* differences in patient risk across physicians. This assumption is unlikely to hold because there is rich variation in patient risk across doctors which is difficult to observe in claims data. For example, some doctors see more patients who report chest pain and are thus more likely to have a pulmonary embolis conditional on observables. For this reason, we augment the model by allowing the probability of a positive test to vary across doctors, conditional on observed patient characteristics.

In particular, we assume that the probability of a positive test is given by:

$$q_{id} = x_{id}\beta + \alpha_d + \eta_{id} \tag{2}$$

where x_{id} are observed patient characteristics, 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. With the introduction of α_d, we now allow patients' risk for PE to vary systematically across doctors. So far, we have modeled two of the three channels that may drive differences in testing behavior: reason 1 (physician practice style) and reason 2 (heterogeneity in PE risk).

In this model, $P(test) = f(x_{id}\beta + \alpha_d - \tau_d - c)$, so the testing equation is only sufficient to identify $\theta_d = \alpha_d - \tau_d$. If we only analyze the probability of testing, we cannot tell if a given doctor tests a lot given observables because she is an overtester (small τ_d) or because she has a patient population which is particularly predisposed to pulmonary embolism (large α_d).

We can distinguish α_d and τ_d if we also observe the number of *positive* tests for each doctor. If a doctor tests more patients given observables because she has a large α_d, i.e. a riskier patient population, then she should also produce more positive tests. So we can separately identify α_d and τ_d by examining the frequency of positive tests conditional on the observed propensity to test.

For example, imagine that doctor A and doctor B have observably similar patients and both test the same fraction of patients. The observed similarity in testing behavior might arise because doctors have comparable testing preferences, i.e. the same testing threshold τ_d; in this case, both doctors must also see patients with similar PE risk, i.e. similar α_d's. With identical thresholds τ_d and patient risk α_d, we would observe not only the same fraction of patients tested, but also the same rates of positive test results amongst tested patients of both doctors.

Alternatively, it might be that doctor A has a lower testing threshold than doctor B ($\tau_A < \tau_B$) but doctor A also has a patient population at lower risk for PE (i.e. $\alpha_A < \alpha_b$), so that on net, doctor A has the same propensity to test as doctor B. In this latter case, we should observe fewer positive tests for doctor A. As we shall see in equation 6 below, controlling for the propensity to test, a doctor with fewer positive tests must have a lower threshold τ_d; only a doctor with a lower testing threshold would test the same fraction of patients despite treating a lower risk patient population. Thus, by using information on both testing behavior and testing outcomes, we can separately identify heterogeneity in PE risk from heterogeneity in testing thresholds.

4.2 Full Model with Rational Weighting

Let us now lay out a more complete version of the model which also allows for heterogeneity in patient benefits of treatment (reason 3) and greater heterogeneity in unobserved patient characteristics (allowing the variance to vary as well as the mean) . Recall our assumption that negative tests are not *ex post* medically valuable; thus, the net benefits of testing are given by the doctor's perceived probability of a positive test (q_{id}) times the net utility of treating a patient who has tested positive, NU_{id}, minus the cost of testing, c.[7] Together, these assumptions imply that doctor d tests patient i if and only if:

$$q_{id}NU_{id} - c \geq \tau_d \tag{3}$$

Define $\theta_d = NU_{id}\alpha_d - \tau_d$. Plugging our specification for the probability of a positive test from equation 2 into the testing equation and dividing through by NU_{id} yields the final form of the testing equation:

$$x_{id}\beta + \frac{\theta_{id} - c}{NU_{id}} + \eta_{id} \geq 0 \tag{4}$$

These assumptions yield a binary choice model of testing. We will assume below that cost c is observable and that net utility of treatment NU_{id} is either known or varies with observables according to a linear function that we estimate. We consider several distributional assumptions about the degree of heteroskedasticity in η_{id} which are described in Section 5.1.

We next show how the testing threshold parameters τ_d can be recovered from a regression of the frequency of positive tests on doctor fixed effects controlling for the propensity estimated from the testing equation. We denote this testing propensity by $I_{id} \equiv x_{id}\beta + \frac{\theta_{id}-c}{NU_{id}}$. Note that net benefits are given by $B_{id} = q_{id}NU_{id} - c = NU_{id}I_{id} + \tau_d + \eta_{id}$. Thus,

$$E(B_{id}|T_{id} = 1) = \tau_d + NU_{id}I_{id} + NU_{id}g_q(I_{id}) \tag{5}$$

where $g_q(I_{id}) = E(\eta_{id}| - \eta_{id} \leq I_{id})$ is an (unknown) function of I_{id}. If we relax the homoskedacity assumption described in the previous section and allow the distribution of η to vary across subsets of the population denoted by q, the function g will also need to be separately estimated for each of those subsets.

Let Z_{id} be an indicator for whether a test was positive or negative. If doctors have rational expectations, we must have $E(q_{id}|T_{id} = 1) = E(Z_{id}|T_{id} = 1)$.[8] Given these rational expectations and equation 3, we can write the expected benefits as $E(B_{id}|T_{id} = 1) = NU_{id}E(Z_{id}|T_{id} = 1) - c$. Plugging this into equation 5 and rearranging yields:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c}{NU_{id}} + I_{id} + g_q(I_{id}) \tag{6}$$

This equation implies that we can recover the testing thresholds τ_d (relative to a normalization) from a regression of the observed testing outcome (positive or negative) on doctor fixed effects, controlling for the estimated propensity to test I_{id}.

[7] The model and estimation generalize to allow for observable patient-level heterogeneity in the cost of testing, but we have calibrated the model with a constant cost c in our estimation

[8] In Section 4.3, we relax this assumption.

In principal, the function $g_q(.)$ may include a constant which would mean that τ_d could only be identified up to a normalization. In order to recover the absolute magnitude of τ_d, we can use the fact that for marginal patients, the expected benefits of testing are exactly equal to the threshold parameter τ_d.

Formally, note that η_{id} is bounded since $q_{id} \in [0, 1]$. Thus, there exists a value $\underline{I}$ such that, for $I_{id} < \underline{I}$, patient i will never be tested. Further, if we observe a tested patient for whom $I_{id} = \underline{I}$, we know that $\eta_{id} = \overline{\eta}$. In other words,

$$\lim_{I \to \underline{I}} g_q(I_{id}) = \lim_{I \to \underline{I}} E(\eta_{id} | \eta_{id} \geq -\underline{I}) = -\underline{I} \tag{7}$$

From, equation 6, this implies that: $E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c}{NU_{id}}$ among patients with $I_{id} = \underline{I}$. To implement this constraint in the model, it is sufficient to normalize I_{id} within each group q so that $I_{id} + g_q(I_{id}) = 0$. We can do this by subtracting a (group q-specific) constant from I_{id} so that $min_{i \in q} I_{id} = 0$ and omitting the constant from estimation of the polynomial $I_{id} + g_q(I_{id})$. This procedure estimates an arbitrary polynomial in I_{id} for each group with an intercept of 0 at the group-specific minimum of I_{id}.

Figure 2 clarifies the intuition for the model and suggests a specification check we can use to evaluate the impact of distributional assumptions about η_{id}. The figure shows a (fictional) relationship between the propensity to test I_{id} and the net benefits of testing for several doctors. For marginal patients with I_{id} close to the minimum value for their group (which is normalized to 0), net benefits are exactly equal to τ_d; the marginal patient reveals the threshold probability at which doctors are willing to test. For the average patient tested, net benefits exceed τ_d and they will exceed τ_d by more at higher propensities since a greater share of patients will be inframarginal with net benefits far exceeding the threshold parameter τ_d.

For doctors within a group (i.e. for doctors with the same distribution of η_{id}), the polynomial relating net benefits to I_{id} is the same for all doctors and only the y-intercept—τ_d—differs. Suppose there is heteroskedasticity such that doctors in group 2 have a higher variance of η_{id}, and thus more private information than doctors in group 1. In that case, the relationship between expected benefits and the propensity to test will be *less* steeply sloped in group 2 because observable characteristics will predict less of the variation in observed rates of positive tests.

Making appropriate distributional assumptions about η_{id} is thus necessary in order to use non-marginal patients to help identify τ_d. This suggests a specification check. We can restrict to doctors with marginal patients and identify the distribution of τ_d by looking only at those marginal patients. Next, we can re-estimate τ_d for those same doctors using the full sample of patients. If the full sample estimates are comparable to the estimates derived using only marginal patients, this suggests that our results are not driven by the functional form assumptions about η_{id} which allow us to infer τ_d using non-marginal patients and the shape of the function $g_q(.)$. We perform this specification check in section 5.2 below.

4.3 Full Model with Misweighting

The model above also assumes that the only "mistake" doctors can make is to test patients if the costs exceed the benefits.[9]

In this section, we consider the additional possibility that doctors systematically misweight observable risk factors in selecting which patients to test. If two patient risk factors predict an equal increase in the probability that a physician orders a CT scan, then these two factors should also predict an equal change in the probability of a positive test after controlling appropriately for unobservables and heterogeneity in net benefits. We identify misweighting by investigating deviations from this prediction—for example, by identifying risk factors that have a strong influence on the probability of a positive test, but little impact on physicians' testing choices. We formalize this intuition below.

First, we relax our assumption that doctors have rational expectations, and assume that each doctor's belief about the probability of a positive test is given by:

$$q'_{id} = x_{id}\beta' + \alpha'_d + \eta_{id} \tag{8}$$

while the actual probability remains:

$$q_{id} = x_{id}\beta + \alpha_d + \eta_{id} \tag{9}$$

With this change, the derivation of the model in Section 4.2 continues to hold, except for the rational expectations assumption. In particular, it is no longer the case that :

$$E(q'_{id}|T_{id} = 1) = E(q_{id}|T_{id} = 1) = E(Z_{id}|T_{id} = 1) \tag{10}$$

Instead we have:

$$E(Z_{id}|T_{id} = 1) = E(q'_{id}|T_{id} = 1) + x_{id}(\beta - \beta') + \alpha_d - \alpha'_d \tag{11}$$

which yields the equation:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c}{NU_{id}} + x_{id}(\beta - \beta') + \alpha_d - \alpha'_d + I_{id} + g_q(I_{id}) \tag{12}$$

While τ_d and $\alpha - \alpha_d$ are in principle separately identified provided $\delta \neq 0$, in practice there is insufficient variation to estimate both. To go further, we make the conservative assumption that doctors misweight observables but are correct on average about the probability that their patients will have PEs. In this case, $E_d(q'_{id}) = E_d(q_{id})$, which implies $E_d(x_{id})\beta' + \alpha'_d = E_d(x_{id})\beta + \alpha_d$. Rearranging, we obtain, $\alpha_d - \alpha'_d = E_d(x_{id})(\beta' - \beta)$. Plugging into equation 12, we obtain:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c}{NU_{id}} + (x_{id} - E_d(x_{id}))(\beta - \beta') + I_{id} + g_q(I_{id}) \tag{13}$$

This equation allows us to recover τ_d *and* $\beta - \beta'$. So we can test both whether doctors have

[9] Testing a patient with negative expected net benefits is an error from the perspective of maximizing social welfare, but perhaps not from the perspective of the doctor if he has a different objective function.

lower testing thresholds or test the wrong patients because they misweight patient characteristics. Intuitively, non-zero coefficients on the (demeaned) x's in equation 12 imply that the x's still have explanatory power in predicting positive tests even after conditioning on doctors' decisions of whether or not to test. Using this model we can simulate how welfare would change if doctors appropriately weighted observables in deciding which patients to test.

To report our results, we proceed in two stages. First, we show the distribution of τ_d and the extent of misweighting estimated in a "positive" model in which we assume that $NU_{id} = 1$ and $c = 0$. This exercise corresponds closely to our reduced form results in that we are trying to document the degree of heterogeneity in testing behavior; our new structural estimates have the additional feature that they allow for doctor-specific unobservable heterogeneity in patient PE risk and can identify misweighting. In this positive model, we take our estimated τ_d and regress them on doctor, hospital and regional characteristics in order to investigate the determinants of heterogeneity in testing behavior.

Second, we discuss calibration of NU_{id} and c and present estimates of "normative" τ_d expressed in dollar terms, the parameter that tells us whether doctors are under or overtesting. We use this model to simulate how testing behavior would differ in a world with no overtesting and in a world in which doctors optimally weighted patient characteristics.

5 Estimation and Calibration of the Structural Model

5.1 Estimation

Equation 4 defines a semiparametric binary choice model which we estimate using Klein and Spady's binary choice estimator (Klein and Spady (1993)). Let t_{id} denote the indicator for whether patient i was tested and let g denote the probability that patient i is tested given index $X_i'\beta$. The log likelihood is given by:

$$L(\beta, g) = \sum_i [t_i \ln g(X_i'\beta) + (1 - t_i)(1 - \ln g(X_i'\beta))] \tag{14}$$

The idea of the Klein-Spady estimator is to approximate g using a "leave-one-out" estimator which predicts the probability of testing for a given patient giving more weight to patients with nearby indices $I_{id} = X_{id}'\beta$. Specifically, we substitute for g using:

$$\hat{g}_{-i,d} = \frac{\sum_{j \neq i} k\left(\frac{(X_j - X_i)'\beta}{h}\right) t_i}{\sum_{j \neq i} k\left(\frac{(X_j - X_i)'\beta}{h}\right)} \tag{15}$$

We use a 4th-order Gaussian Kernel and empirically select for the smallest bandwidth such that $\hat{g}$ is a monotonic function of the index $X_i'\beta$.

This method of estimation implicitly assumes that η_{id} in equation 4 are i.i.d. across patients and doctors within quantiles. There are several reasons this assumption might be violated. One concern is that a doctor who tests 5% of patients sees a fundamentally different population of patients from a doctor who tests only 0.5% of patients; for these doctors not only does the average population PE

risk α_d differ, but also the distribution of unobserved (to the econometrician) PE risk η_{id} may differ. Perhaps the doctor who tests 5% of patients sees many patients whose (unobservable) symptoms are marginal for pulmonary embolism while the doctor who tests 0.5% of patients is more likely to see patients with either no chance of a pulmonary embolism or a very high chance, corresponding to a higher variance of η_{id} for the second doctor.

To deal with these concerns, we split the population of physicians into 10 deciles based on the proportion of each physician's patients tested. The testing model is estimated separately within each of those deciles, so that the assumption of homoskedastic η_{id} is only imposed within deciles. This way, we are comparing doctors who tested 5% of patients to doctors who tested 5.5%, but we are not comparing doctors who tested 5% to doctors who tested 0.5%.

A further concern is that doctors may have different η distributions because they differ in their ability to accurately diagnose a pulmonary embolism given the available data. We consider an explicit model of this in Section 6.2 where we consider a parametric model in which η_{id}'s distribution can be flexibly estimated for each doctor.

Given the propensity to test I_{id} from estimating equation 4, the next step is to estimate the net benefit equation:

$$E(Z_{id}|T_{id}=1) = \frac{\tau_d + c}{NU_{id}} + (x_{id} - E_d(x_{id}))(\beta - \beta') + \tau_d + h_q(I_{id}) \tag{16}$$

where $h_q(I_{id}) = I_{id} + g_q(I_{id})$. We estimate this equation by OLS using a cubic polynomial to estimate each of the 10 $h_q(.)$ functions.

The distribution of estimated $\hat{\tau}_d$ has larger variance than the underlying distribution of true τ_d due to estimation error. To correct for this, we construct "empirical Bayes" estimates of τ_d. More precisely, we show in Appendix A how to recover the mean and variance of τ_d from the estimated distribution of $\hat{\tau}_d$; these moments are identified without any additional distributional assumption.

5.2 Estimates with $NU_{id} = 1$ and $c = 0$

We start by estimating a simplified version of the model in which net benefits are set to a constant value of one, $NU_{id} = 1$, and there are assumed to be no costs of testing, $c = 0$. Under these simplifying assumptions, estimation of the net benefit equation (equation 13) echos the reduced form analysis in Section 3.5, where we regressed an indicator variable for a positive test on a vector of patient characteristics. The main difference is that here we control for the propensity to test recovered from estimation of the testing equation (equation 4). Assuming the model is correctly specified, this allows us to control for systematic differences in patient mix (different values of α_d) across doctors. Assuming no heterogeneity in c or NU_{id}, the estimated τ_d can be interpreted as the structural cut-off probability of a positive test at which doctors would be willing to test a marginal patient.

5.2.1 Distribution of τ_d and Misweighting

We begin by estimating the testing equation, equation 4. For each of the deciles defined above, we estimate a separate testing equation. Figure 3 shows the relationship between the underlying

estimated propensity to be tested, I_{id} and the observed probability of testing for a single group q (the propensity to test is normalized separately within each group and cannot be compared across groups).

As we might expect, this function is convex for large values: for most patients a single risk factor is not worrying, but in the presence of several other warning signs the marginal impact on the likelihood of testing increases. Note also that this function can be approximated well by a piecewise linear function. In the parametric version of the model we consider below, the critical simplifying assumption will be that this function must be piecewise linear (which corresponds to an underlying uniform distribution of η_{id}).

Next we consider the results from estimation of the net benefit equation. Figure 4 gives a graphical check of the estimation procedure. It is the empirical analogue of the illustrative figure 2. In blue, we plot the fraction of tests which are positive along the y-axis against the estimated propensity to test (I_{id}) along the x-axis. We are plotting this relationship within a single decile of physicians, i.e. within a single group where η_{id} is homoskedastic across doctors. As expected, we see that the proportion of positive tests is increasing for patients with a higher predicted probability of being tested. In other words, the observable characteristics that predict a high probability of being tested also predict a higher probability of a positive test, amongst tested patients.

Note that the model's predicted probability of a positive test given the estimated testing propensity I_{id} (shown in red) is a good fit for the realized probability of a positive test (shown in blue). To simplify the visual presentation of this graph and highlight the estimated mapping of testing propensity to positive tests, we have averaged across doctors, each of whom has a different y-intercept τ_d as illustrated in figure 2. If we were to graph each doctor separately, we would see a common polynomial fit with varying intercepts across doctors corresponding to the estimated thresholds τ_d. We conclude that the model of positive tests, which fits a single polynomial of the testing propensity I_{id} plus doctor-specific intercepts, fits the observed data well as indicated by the close coincidence of red and blue data points.

The distribution of estimated $\hat{\tau}_d$ 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 this, we consider empirical Bayes estimates in order to recover moments of the true underlying distribution of τ_d. Our approach is described in detail in Appendix A. Unlike more standard estimators (Kane and Staiger 2008), our approach 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. Here, we estimate the mean and the standard deviation.

The mean value of τ_d is .039 and and the standard deviation is 0.040.[10] 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 3.9%. Note that this includes tests which detect actual pulmonary embolisms and false positives, so a 3.9% probability of a positive test is *not* the same as a 3.9% probability of a pulmonary embolism.

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

The standard deviation of τ_d recovered here, 0.04, is very close to the standard deviation of the doctor fixed effects estimated in the reduced form specification, as reported in Section 3.5, also 0.04. This does not mean that unobserved characteristics are unimportant in determining doctors' testing behavior; rather, any upward bias in the reduced form standard deviation due to the measured standard deviation including heterogeneity from (uncontrolled for) unobserved characteristics is apparently offset by downward bias generated by the correlation of these unobserved factors with τ_d.

The standard deviation of the misweighting term $(x_{id} - E_d(x_{id}))(\beta - \beta')$ is in fact larger than the standard deviation of τ_d: the average absolute value of the misweighting is .030 and the standard deviation is .050. This is an extremely large effect: the average tested patient has a 7% probability of having a pulmonary embolism and doctors misperceive that probability by an average of 3% due to improperly weighting observables.

Table 2 breaks down further the sources of the misweighting. The first column shows the average marginal effects in the structural model. The bias term reports our estimate of $\beta - \beta'$. We find evidence of significant misweighting for 16 of the 37 variables considered in our testing model. Our largest effect is for a history of pulmonary circulation disease; these patients are 28 percentage points more likely to have a pulmonary embolism conditional on being tested (almost 6 times as likely), but our model suggests that doctors are no more likely to test them than other patients.

Doctors also give insufficient weight to whether a patient was recently admitted to the hospital in deciding whether to test them; overall, doctors are no more likely to test these patients although they are about 10 percentage points more likely than other patients to have pulmonary embolisms. We find a similar pattern for metastatic cancer, coagulation deficiency, obesity, and neurological disorders. Doctors appear to give slightly too much weight to peripheral vascular disease in deciding which patients to test.

In order to test the structural assumptions of our model, we next consider what values of τ_d we would estimate using only marginal patients. The basic idea is as follows: using the propensity from our testing equation, we can identify patients who are marginal in the sense that they are just barely worth testing; these are patients with the lowest propensity conditional on being tested. For doctors with marginal patients, we can directly compute τ_d by just observing the rate of positive tests among marginal patients without needing to estimate equation 16.

To test whether the full model is correctly specified, we can restrict to doctors who have some marginal patients and ask whether the mean values of τ_d estimated in the full model from equation 16 agree with the values computed directly from marginal patients. Specifically, we define marginal patients as patients in the first quintile of the estimated propensity to test within each patient decile (the propensities cannot be compared across deciles). This test finds very close agreement between the marginal estimates and the full model estimates: the mean of the full model estimates among doctors with marginal patients is 0.0426 and the mean of the direct estimates is 0.0419. The direct estimate τ_d relies only on the assumption that we have succeeded in identifying patients who are marginal given observables; the estimates in our full structural model require additionally that we have correctly estimated the function which relates the probability of testing to the propensity to test for non-marginal patients. This specification test provides some reassurance that that function

is correctly specified.

The estimated values of τ_d are hard to interpret in isolation. Below we consider how far testing deviates from the efficient benchmark given various assumptions about the cost of testing, the likelihood of false positives, and the net utility conditional on finding a positive test. Additionally, some of the measured heterogeneity in τ_d could reflect heterogeneity in the net utility conditional on a positive test or heterogeneity across doctors within deciles in the distribution of η. We consider versions of our model which allow for both of these forms of heterogeneity below.

5.2.2 Determinants of τ_d

We next consider regressions of the estimated τ_d on doctor, hospital and regional characteristics in order to better understand the determinants of doctor practice style. Specifically, we regress τ_d on variables capturing doctor experience (the number of years since the doctor graduated from medical school), whether the medical school the doctor attended is ranked in the top 50 for research by US News & World Report, whether the medical school is ranked in the top 50 for primary care, whether the hospital where the scan was performed is a for profit hospital or an cademic hospital, regional medical spending, and average income in the region where the hospital is located.

We consider OLS estimates as well as FGLS estimates which take into account the estimation error in the dependent variable τ_d.[11] For each estimate, we consider models with and without hospital fixed effects. Including hospital fixed effects to identify the impact of within-hospital variation in physician characteristics obviates the concern that our model omits unobserved differences in the cost of testing at the hospital level. For example, there may be variation in the opportunity cost of testing, depending on whether the CT scan is used to capacity. This heterogeneity will be absorbed into the hospital fixed effect.

We also consider models in which we estimate a separate threshold parameter τ_d for each doctor in each year, allowing for time variation in doctor fixed effects. We then run a panel data regression including both doctor and year fixed effects, allowing us to separate the impact of physician experience from cohort effects. These models are estimated on a more restrictive sample since we keep only those doctor-year pairs for which there are at least 3 observed tests.

Table 3 reports the results. Our most consistently robust finding is that doctors in higher spending regions have lower testing thresholds, i.e. they are more likely to test low risk patients. A 10% increase in regional spending correlates with a 0.4 percentage point decline in the testing thresholds. We also find some evidence that more experienced doctors have higher testing thresholds, although only the across doctor variation is sufficiently precise to measure this effect. A 10-year increase in doctor experience is associated with an 0.3-0.4 percentage point increase in the testing threshold depending on the specification (recall that a larger threshold is associated with less testing). Finally, we find some evidence in the cross-sectional specification that doctors practicing in academic hospitals have higher testing thresholds, but this effect is not statistically significant at the 5% level.

[11] The FGLS estimates are based on Lewis and Linzer (2005), where the error term consists of both a homoskedastic ϵ_{id} with unknown variance and a heteroskedastic component with known variance. The heteroskedastic component arises from the estimation error in τ_d which is in turn recovered from estimation of equation 16.

5.3 Estimates of τ_d with Calibrated NU_{id} and c

In this section, we further calibrate the model estimated in the previous section so that we can determine the dollar value of τ and therefore which doctors are overtesters and undertesters. This requires making assumptions about the benefits of treatment conditional on finding a pulmonary embolism, the financial and health costs of testing and treatment, and the likelihood of false positives.

5.3.1 Calibration of Parameters

We need to determine the values of c and NU_{id} for each patient. An important cost of overtesting comes from the fact that tests have both type I and type II errors, so overtesting leads to unnecessary treatment which can have adverse consequences. CT scans, as with many other medical tests, can generate both false positive and false negative results (Stein et al. 2006). In this section, we extend the model to explicitly include false positives and negatives, and then describe our calibration.

Let s denote the sensitivity of the test (one minus the probability of a false negative) and fp denote the probability of a false positive (one minus the specificity). Let PE_{id} denote the event that patient i actually has a PE. As before, Z_{id} is an indicator which is 1 if a test is positive. MB_{id} denotes the medical benefits of treatment if the patient has a PE, Mc denotes the medical costs of treatment and CT_{id} denotes the financial cost of treatment. Then the net utility of a positive test is given by:

$$NU_{id} = P(PE_{id}|Z_{id} = 1)MB_{id} - MC_{id} - CT_{id} \tag{17}$$

The benefits of treatment are incurred only if the positive test result is a "true positive,", i.e. the patient actually has a PE. The medical risks and financial costs of treatment are incurred for any treated patient regardless of whether he actually has a PE.

Applying Bayes' Rule and the law of total probability we can rewrite this in terms of s and fp as:

$$NU_{id} = \frac{s(q_{id} - fp)}{q_{id}(s - fp)}MB_{id} - MC_{id} - CT_{id} \tag{18}$$

We can therefore write the net benefits of testing as:

$$\begin{aligned} B_{id} &= q_{id}NU_{id} - c \\ &= \frac{s(q_{id} - fp)}{(s - fp)}MB_{id} - q_{id}MC_{id} - q_{id}CT_{id} - c \end{aligned} \tag{19}$$

Let $\hat{NU}_{id} = \frac{s}{s-fp}MB_{id} - MC_{id} - CT_{id}$ and $\hat{c}_{id} = c + \frac{s \cdot fp}{s-fp}MB_{id}$. Then we can rewrite the net benefits of testing as:

$$B_{id} = q_{id}\hat{NU}_{id} - \hat{c}_{id} \tag{20}$$

which is exactly the definition of net benefits in Section 4.

Conditional on whether or not testing and treatment are observed, false positives and false negatives impact only marginal benefits. False positives and false negatives do not impact the costs of testing, since those are paid if a test is done regardless of outcome, and similarly, they do not affect the costs of treatment, which are paid if treatment is performed. On the other hand, the

benefits of treatment accrue only if the patient actually has the underlying condition; a patient with a false positive experiences no benefit from treatment. If there are more false positives, the marginal benefits of any observed positive test will be smaller.

We calibrate these parameters using the values in Table 4, based on recent numbers from the medical literature on pulmonary embolism. Note that our calibration of both the medical benefits and the medical cost of treatment depend on an estimate of the value of a statistical life (VSL). To the extent that we use a higher VSL, the cost of treatment and the cost of testing c will be proportionately less important (and so testing will be more desirable). Below, we consider a variety of different fixed values for the VSL. A value between \$1 and \$2 million would be consistent with a \$100,000 value of a statistical life year given the life expectancies observed in our data. In the parametric version of the model, we allow the value of a statistical life to vary with observables such as age.

5.3.2 Distribution of τ_d and misweighting with calibrated parameters

Table 5 shows the estimated values of τ_d given the parameter values assumed in the previous section for different assumptions about the value of a statistical life and the rate of false positives. To determine the percentage of doctors overtesting we need to make a distributional assumption beyond the mean and variance estimated so far. First, note that τ_d is bounded below. As long as NU_{id} is positive (which it is for all values of the calibrated parameters), the minimum possible value of τ_d is achieved when $q_{id} = fp$ so that all positive tests are false positives. At this value, $B_{id} = -fpMC_{id} - fpCT_{id} - c$; there are no medical benefits of testing and the "net benefit" is given by the medical costs of treatment for false positives, the financial costs of treatment for false positives and the costs of testing. We assume that τ_d minus this minimum value is log-normally distributed with the estimated mean and variance. Table 5 additionally reports the percentage of doctors overtesting at each set of parameter values given this distributional assumption.

The table suggests that the degree of overtesting is extremely sensitive to our assumptions about the rate of false positive tests and the value of a statistical life. The existing literature suggests a false positive rate of 0.04 (Stein et al. 2006) which would imply, depending on the VSL, that 98-99% of doctors are overtesting and that doctors test a patient even if the costs of testing exceed the benefits by between \$500 and \$1000 on average.[12] While we cannot directly observe which tests in our data were false positives, this rate is high enough to be problematic because there are some patients for whom, conditional on observables, the proportion of positive tests is less than 0.04. See for example Figure 4 where the marginal patient has between a 2% and 4% likelihood of a positive test. This would be a contradiction if the false positive rate were exactly 0.04 for all patients since such a false positive rate would yield an observed fraction of positive tests of 4% even if no one had pulmonary embolisms.

There are two ways of dealing with this problem. One is to assume that the true false positive rate must be lower than documented in the medical literature, and in Table 5 we consider several

[12]Note that given our assumption of log normal distribution of τ_d thresholds, even when the mean of the distribution if positive, as it is for low values of the false positive rate, we find in some cases that the majority of doctors still are overtesters.

different values. A second approach is to assume that the false positive rate varies across patients. Thus, we also consider a model in which the average false positive rate is 0.04 but the false positive rate within each decile of patients is bounded from above by the observed rate of positive tests among marginal patients.[13] The final column of Table 5 shows these results. The upshot is that it is the average rate which is driving our results on the distribution of τ_d—as long as the average rate of false positives exceeds 0.03, we find that a large majority of doctors are overtesting.

The second part of Table 5 shows the mean absolute deviation and standard deviation of the misweighting term. This value is sensitive to the VSL but not to the false positive rate as we would expect (the stakes of testing scale with the VSL while the false positive rate is–to a first approximation—a location shifter for the distribution of net benefits). The absolute value of the misweighting term is more than twice as large in magnitude as the mean value of τ_d at a false positive rate of 0.03 and still more than 30% larger at a false positive rate of .04.

The welfare consequences of $\tau \neq 0$ and the misweighting term depend on the number of patients for whom doctors change their testing behavior. If all patients should clearly be tested or not tested, the welfare impact of both will be minimal; if there are a large number of marginal patients, the welfare impact will be larger. The next section takes this into account in order to determine the magnitude of the welfare loss due to overtesting and misweighting. But first, we can perform a back of the envelope calculation to assess their relative importance.

Assume for simplicity that the distribution of net benefits is uniform (this is a different uniformity assumption than we make in the model below and is made here only to simplify). At a false positive rate of 0.03, the average absolute value of misweighting is \$650 and the average τ is -\$250. Given τ of -\$250, every patient with net benefits between \$0 and -\$250 will be tested even though they should not be, with an average welfare loss of \$125 (given our simplifying uniformity assumption). Suppose there are N such patients so that the total welfare loss from overtesting is \$125N.

Now consider the welfare loss from misweighting. Assume for the moment that the average absolute misweighting of \$650 reflects a 50% chance of a misweighting of +\$650 and a 50% chance of -\$650. Let's consider patients in [0,\$650].[14] For these patients, there is a 50% chance that the misweighting is -\$650 leading to the wrong decision (no test when they should be tested). The average welfare loss per patient is \$325. The number of impacted patients is $2.6N$ (where N is the number of patients in [0,250]). The total welfare loss from undertesting due to misallocation is thus: $1/2 \cdot 325 \cdot 2.6N \approx 425N$; this is over three times the size of the welfare loss from overtesting due to low thresholds τ_d.

Note that this calculation is conservative with respect to the relative benefits of misweighting: the degree of misweighting has greater variance as well and the welfare loss is convex as a function of the degree of misweighting. If the misweighting of -\$650 is instead a 50% chance of 0 and a 50% chance of \$1300, the welfare loss from undertesting due to misweighting will double.

The next section performs this calculation in a more formal way accounting for the correct

[13] In deciles where this bound is not binding, we use the smallest constant consistent with an overall average of 0.04.

[14] For patients with negative net benefits the calculation is more complicated since misweighting sometimes offsets overtesting due to low τ_d—the net effect of misweighting on welfare for these patients is negative but smaller than the welfare loss from undertesting. The welfare loss from overtesting is bounded from below by the medical and financial costs of testing, which contributes to the much lower welfare loss of misweighting from overtesting compared to the welfare loss of misweighting driven by undertesting.

underlying distribution of net benefits and the heterogeneity in both τ_d and the misweighting term. It also determines the absolute welfare loss in each scenario by using the model to determine the appropriate value of N.

6 Simulations and Welfare

In this section we perform several simulations to determine how welfare would change if doctors behaved optimally from a social standpoint. We begin by simulating worlds with no overtesting or no misweighting and using the model to compute the welfare gain, decomposing this gain into financial and medical costs and benefits. This simulation will require we make parametric assumptions about the distribution of η_d but these parametric assumptions also allow us more leeway to allow net utility to vary with observables. Next, we consider a generalization of our testing model which allows doctors to vary both in their threshold parameters τ_d and in their ability to determine which patients should be tested given unobservables (which we model as heterogeneity in the variance of η_{id} across doctors).

6.1 Welfare Cost of Overtesting and Misweighting

The model implies welfare loss whenever $\tau_d \neq 0$. We focus on the case of overtesting ($\tau_d < 0$) for two reasons. First, overtesting is empirically the larger problem in our sample, with an estimated 95-99% of doctors overtesting, under our preferred calibration assumptions. Second, we find that the welfare loss due to $\tau_d > 0$ is highly dependent on the distribution we assume for τ_d (of which we estimate only the mean and variance); for some distributions even a small number of doctors undertesting can lead to large welfare losses from undertesting if the right-tail is sufficiently thick.

In order to simulate the impact of $\tau_d < 0$ on testing behavior, we must first be more explicit about the link between the net benefit equation (equation 6) and the testing equation (equation 13). Our estimates so far give τ_d in units either of the probability of a positive test (as in section 5.2.1) or in units of dollars, while the propensity in our testing equation is only semiparametrically identified up to a location and scale transformation.[15]

To relate the testing equation to the net benefit equation, we will assume that the distribution of η_{id} is uniform so that shape of the function $g_q(I_{id})$ is known (an assumption we validate below). Given this assumption, we can recover the scaling factor relating the propensities estimated in the testing equation to the treatment equation (and thus, we can determine how changing τ_d would impact testing behavior).

Let $\eta_{id} \sim U(-\eta, \eta)$ and assume that η is such that there are no observables for which doctors always want to test regardless of η_{id}. We show in Appendix B that this implies:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c}{NU_{id}} + \frac{\lambda \hat{I}_{id}}{2} + (x_{id} - E_d(x_{id}))(\beta - \beta') \tag{21}$$

Therefore, the scaling factor λ can be recovered from a regression of function $h_q(I_{id}) = NU_{id}(I_{id} +$

[15] In our setting, there is no obvious price to put into the testing equation in order to express it in dollar terms. The fee c that is paid per test is paid to the radiologist rather than the diagnostician ordering the test.

$g_q(I_{id})) = \frac{\lambda \hat{I}_{id}}{2}$ on the estimated propensity $\hat{I}_{id}$. Using this scaling factor, we can determine how the true propensity would change were τ_d set to 0. Note that the assumption that η_{id} is uniformly distributed is empirically validated by Figure 3. When we estimate $g_q(I_{id})$ in a completely flexible way, we find that it appears piecewise linear, exactly what we expect if η is uniformly distributed.

With the additional power afforded by the assumption that η_{id} is uniform, we can also allow NU_{id} to vary across patients. Specifically, we allow that $NU_{id} = \overline{NU} + x_{id}\delta$. For example, age might impact net utility through several avenues such as the safety of treatment with blood thinners as well as the value of a statistical life.

Using this model, we compute separately the medical benefits of testing, the medical costs of testing, the financial costs of testing and the net benefits of testing given the estimated τ_d and in a hypothetical world where $\tau_d = 0$ for different values of the false positive rate. These results are shown in Table 6. As the false positive rate (and thus the degree of overtesting) increases, the welfare benefits from eliminating overtesting also rise and the overall net benefits of testing fall. For a false positive rate of 3% or 4%, the fraction of patients tested would drop by more than 40%. The financial costs of testing would fall by a proportionate amount and 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 actually having a pulmonary embolism. The net benefits of testing would increase by more than 30% with a false positive rate of .03 and more as the false positive rate increases.

Table 7 simulates a world in which doctors weighted observables in the manner the model suggests is optimal. Properly weighting observables leads a higher fraction of patients to be tested. But by far the predominant factor is the increase in the number of pulmonary embolisms detected. The medical benefits due to treatment of PE increase by more than 90% and net benefits increase by 275%. Undiagnosed PE is thought to be a major public health problem, with the Surgeon General (2008) estimating that approximately half of PE cases are never diagnosed; analysis of autopsy reports have found it to be a frequently missed mortality risk. By improving physician assessment of patient PE risk, our model suggests that the rate of undiagnosed PE could fall substantially.

6.2 Heterogeneity in Doctor Discernment

The model so far has assumed that doctors who test a similar fraction of patients do not differ in their ability to use unobservables to determine which patients should be tested. Specifically, we have assumed that η_{id} is i.i.d. across doctors within each decile of overall testing behavior. We can relax this assumption by considering a parametric version of our model in which the distribution of η_{id} is allowed to vary across doctors.

We model variation in doctor discernment as heterogeneity in the variance of η_{id} across doctors. Intuitively, some doctors may have no additional information beyond what is observable to the econometrician. For these doctors, $Var(\eta_{id}) = 0$. A higher variance of η_{id} implies that doctors can do a better job of distinguishing between patients who look identical given observables. While one doctor sees several patients with similar observables and assumes they all have a 4% chance of testing positive, a second doctor realizes that some of the patients have a much lower chance while others have a much higher chance (our model so far has allowed for private information but assumed

that it is identically distributed across doctors).

The model now captures a key trade-off observed by Doyle, Ewer, and Wagner (2010). They find in a natural experiment that physicians from more prestigious residency programs achieve similar patient outcomes at 10-25% lower cost. One explanation for this phenomenon is that physicians from less prestigious schools administer more wasteful care and could achieve the same outcomes at lower cost if they cut back some services. In the language of our model, these physicians might be overtesting in the sense that they have smaller τ_d. A second explanation is that these physicians just need to use more medical resources to achieve the same quality of care. Perhaps these physicians have the same τ_d as physicians from more prestigious medical schools but they need to test more to achieve the same number of positive tests. In our model, this would appear as a lower variance of η_{id}. We refer to this hypothesis as heterogeneity in doctor discernment.

The model we develop in this section allows us to disentangle these hypotheses with one important caveat. Differences in the distribution of η_{id} across doctors could reflect either differences in doctor behavior for identical patients or differences in the distribution of patient unobservables. Some doctors might see a patient pool where more information is available about whether or not they are likely to have a pulmonary embolism, e.g. due to variation in access to an electronic medical record. Because of this, we cannot necessarily interpret heterogeneity in the measured variance of η_{id} across doctors as evidence of heterogeneity in doctor discernment. Nonetheless, to the extent that differences in τ_d in the previous section were driven by differences in discernment *or* differences in the distribution of patient unobservables available to the doctor, the model in this section separately identifies these factors from heterogeneity in overtesting. This distinction is critical, because it allows us to answer the question of whether welfare would actually improve if those doctors who appear to overtest (in the sense of having fewer positive tests for a given propensity) tested less.

In order to implement this version of the model, we need to make new parametric assumptions about the distribution of η_{id}. Formally, we assume that $\eta_{id} \sim U(-\eta_d, \eta_d)$ and assume that η_d is such that there are no observables for which doctors always want to test regardless of η_{id}. We show in Appendix B that this implies:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c}{NU_{id}} + \frac{I_{id} + \eta_d}{2} \tag{22}$$

Immediately the identification problem from the model in the previous section is clear. The doctor fixed effects that we estimated conditional on the propensity to test could reflect either actual heterogeneity in τ_d (overtesting) or heterogeneity in the variance of η_{id} (differences in doctor discernment).

However, we can now solve this identification problem by directly estimating η_d for each doctor from the testing equation. Intuitively, heteroskedasticity in η_d is identified by the fact that observables are less predictive of testing behavior for doctors with more private information. Given the estimated testing propensities $\hat{I}_{id}$ and $\hat{\eta}_d$, we still need to recover a scaling factor to simulate the impact of differences in τ_d on testing behavior so the net benefit equation we estimate is:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c}{NU_{id}} + \frac{\lambda(\hat{I}_{id} + \hat{\eta}_d)}{2} \tag{23}$$

where now τ_d can be estimated simultaneously with the scaling factor λ in a least squares regression. Note that this version of the model omits the misweighting term; while η_d, τ_d, λ and misweighting are in principle separately identified, we find in practice that λ—and thus the implied τ_d are difficult to estimate precisely once both $\eta - d$ and the misweighting term are included.[16]

The model with heterogeneous η_d is quite similar to the model of cesarean section births studied by Currie and MacLeod (2013). In both cases, the relevance of unobservable factors to the testing/treatment decision is identified based on the slope of the testing/treatment decision with respect to observables. Two important differences are worth noting: first, Currie and MacLeod (2013) argue in their setting that heterogeneity in η_d reflects diagnostic skill and not differences in the patient population; we do not think this assumption is supportable in our setting and so we cannot separately identify doctors with more diagnostic skill of this type. Second, in our setting (testing, as opposed to treatment) a large η_d implies that a doctor knows with high precision whether patients will have a pulmonary embolism—η_d reflects additional information which should be used in deciding whether to test. In contrast, Currie and MacLeod (2013) interpret a large η_d as evidence that doctors have a noisier signal of the true underlying patient benefit from intervention—the best doctors respond only to the index of observables. The testing equation alone cannot tell us which of these assumptions is right. However, the net benefit equation reveals that in our setting, η_d does contain information relevant to forecasting whether tests will be positive because $\lambda > 0$ even after controlling for observables in the net benefit equation (equation 22).

Table 8 shows the results of this analysis. We report the estimated mean and standard deviation of τ_d for different values of the false positive rate and we compare these values to the homoskedastic model estimated under the uniform and semiparametric assumptions. The upshot of this analysis is that ignoring heterogeneity in doctor discretion causes us to slightly understate the degree of overtesting—after allowing for heteroskedasticity in η_{id}, τ_d is \$50-\$150 smaller depending on the specification. The results thus suggest that patient welfare would improve if doctors were more reluctant to test marginal patients, and heterogeneity in testing thresholds was not primarily driven by some doctors with poor discernment needing to test more to find the same number of pulmonary embolisms.

7 Discussion & Conclusion

While it is commonly believed that the US health care system includes significant wasted 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 (allocative inefficiency) and mistargeting of medical resources (productive inefficiency). This paper investigates both forms of inefficiency, analyzing whether doctors efficiently select patients for medical testing and whether they set optimal risk thresholds at which to test patients. We study these inefficiencies in the context of emergency department CT scans to diagnose pulmonary embolism. Our results suggest that CT scans for pulmonary embolism are used more widely than our model suggests would

[16]This is because the term $\hat{I}_{id} + \hat{\eta}_d$ has much higher variance than the corresponding $\hat{I}_{id}$ in the homoskedastic model. If we do include the misweighting term in this regression, the coefficients on the misweighting terms are very similar to those in the previous section but λ - and thus our estimates of τ_d - are extremely imprecise.

be optimal, and that the highest risk patients are not necessarily the ones receiving these scans.

Controlling for relevant differences in the patient population, our model shows that there remains enormous heterogeneity in the net benefit threshold at which doctors will test a patient. Less experienced physicians and those practicing in high-spending regions (as measured by the Dartmouth Atlas) test more patients on average, holding fixed the patients' risk of PE. After making additional calibration assumptions regarding the benefits of treatment conditional on finding a pulmonary embolism and the cost of testing, we estimate that 90-99% of doctors evaluating emergency department patients are performing too many tests, i.e. they are testing patients for whom the medical risks and financial costs of the test exceed the expected medical benefits of treatment. If all doctors tested only when expected benefits exceed expected costs, 40-50% fewer chest CT scans would be performed and the net benefit of CT tests for pulmonary embolism would increase by 20-35% assuming false positive rates estimated in the medical literature.

More surprisingly—despite the fact that we studied this test in part due to anecdotal evidence of systematic overuse—we find that testing the wrong patients yields greater welfare loss than testing too many patients. Doctors systematically underweight certain important predictors of PE risk, including recent prior hospitalizations, obesity, and metastatic cancer. These mistakes 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 (such as their failure to explicitly consider obesity as a risk factor) or failures to make adequate use of these scores. These mistakes in assessing patient PE risk lead to significant welfare losses from failing to test particular high risk patients; these losses are an order of magnitude larger than the welfare loss from overtesting. While physicians test many patients on the margin who should not be tested, they also fail to test patients who, given their observable risk factors, stand to gain a great deal from being tested.

Our estimates are calculated on a 20% sample of patients enrolled in Medicare Parts A and B, and the welfare numbers reported in the paper reflect potential gains to this sample only. To understand the total national welfare loss associated with the inefficiencies we identify in this sample, we do an informal scaling exercise. First, scaling the estimates up by a factor 5, we have estimates that cover the entire population of traditional Medicare enrollees; this corresponds to a $25 million loss from overuse of PE CT due to low testing thresholds, and a $211 million loss from misweighting observable patient risk factors. These numbers would cover an estimated 43,000 diagnosed cases of PE each year amongst traditional Medicare enrollees.

National estimates put total incidence of diagnosed PE in the United States at 350,000 per year (Office of the Surgeon General 2008); about half of these occur amongst patients who are already admitted to the hospital (for other reasons) or in a nursing home (Fagan 2010). Our estimates apply only to the estimated 175,000 PEs that occur amongst outpatients, since we have not modeled the decision to test admitted patients for PE. [17] Assuming that these welfare costs scale with the number

[17] If we assume our estimated rate of PE extends to Medicare Advantage enrollees, who account for 28% of total Medicare patients, then our sample scales to 53,000 total diagnosed PEs occurring outside the hospital in all Medicare patients. This would suggest that about roughly one third of the estimated 150,000 PEs occur in patients over the age of 65, which is consistent with the existing literature's estimates of PE incidence by age (Silverstein et al. 1998).

of diagnosed PEs amongst outpatients to cover the entire population, we estimate $100 million lost from overuse of PE CT, and $840 million lost from misweighting observable patient risk factors at our preferred calibration parameters.

These findings suggest that both overuse and misuse of medical resources are important drivers of high spending and low medical returns to care. By measuring physician-level preferences for under- or over-testing, we are able to explore the training and environmental factors that contribute to overuse. Future work could pair this framework for estimating the overuse of diagnostic testing with experimental or quasi-experimental variation in physician's training or practice environment; together, 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. Our findings also 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

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

Coco, A. S. and D. T. O'Gurek (2012, January-February). 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, M. M., G. Randall, M. Gosselin, M. Brandt, K. Spinning, and C. D. Vegas (2008, August). Ct angiography in the evaluation of acute pulmonary embolus. *American Journal of Roentgenology 191*(2), 471–474.

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

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

Doyle, J. J., S. M. Ewer, and T. H. Wagner (2010). Returns to physician human capital: Evidence from patients randomized to physician teams. *Journal of health economics 29*(6), 866–882.

Elixhauser, A., C. Steiner, D. Harris, and R. Coffey (1998). Comorbidity measures for use with administrative data. *Medical Care 36*(1), 8–27.

Fagan, K. A. (2010). Pulmonary embolism. In D. E. Schraufnagel (Ed.), *Breathing in America: Diseases, Progress, and Hope*, Chapter Pulmonary Embolism, pp. 165–174. American Thoracic Society.

Garber, A. M. and J. Skinner (2008). Is american health care uniquely inefficient? Technical report, National Bureau of Economic Research.

Goldhaber, S. Z. and H. Bounameaux (2012). Pulmonary embolism and deep vein thrombosis. *The Lancet 379*(9828), 1835–1846.

Kane, T. J. and D. O. Staiger (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

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

Lessler, A. L., J. A. Isserman, R. Agarwal, H. I. Palevsky, and J. M. Pines (2010, April). Testing low-risk patients for suspected pulmonary embolism: A decision analysis. *Annals of Emergency Medicine 55*(4), 316–326.

Lewis, J. B. and D. A. Linzer (2005). Estimating regression models in which the dependent variable is based on estimates. *Political Analysis 13*(4), 345–364.

Mamlouk, M. D., E. vanSonnenberg, R. Gosalia, D. Drachman, D. Gridley, J. G. Zamora, G. Casola, and S. Ornstein (2010, August). Pulmonary embolism at ct angiography: Implications for appropriateness, cost, and radiation exposure in 2003 patients. *Radiology 256*, 625–632.

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

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

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

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

Rahimtoola, A. and J. D. Bergin (2005, February). Acute pulmonary embolism: An update on diagnosis and management. *Current Problems in Cardiology 30*, 61–114.

Silverstein, M. D., J. A. Heit, D. N. Mohr, T. M. Petterson, W. M. O'Fallon, and L. J. M. III (1998, March 23). Trends in the incidence of deep vein thrombosis and pulmonary embolism. *Archives of Internal Medicine 158*(6), 585–593.

Stein, P. D., S. E. Fowler, L. R. Goodman, A. Gottschalk, C. A. Hales, R. D. Hull, J. Kenneth V. Leeper, J. John Popovich, D. A. Quinn, T. A. Sos, H. D. Sostman, V. F. Tapson, T. W. Wakefield, J. G. Weg, and P. K. Woodard (2006, June 1). Multidetector computed tomography for acute pulmonary embolism. *New England Journal of Medicine 354*(22), 2317–27.

Venkatesh, A., J. A. Kline, and C. Kabrhel (2013, Jan. 28). Computed tomography in the emergency department setting–reply. *Journal of the American Medical Association Internal Medicine 173*(2), 167–168.

Venkatesh, A. K., J. A. Kline, D. M. Courtney, C. A. C. Jr, M. C. Plewa, K. E. Nordenholz, C. L. Moore, P. B. Richman, H. A. Smithline, D. M. Beam, and C. Kabrhel (2012, July 9). 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), 1028–1032.

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

Table 1: Summary Statistics

	A. Untested patients	*B. Patients with negative tests*	*C. Patients with positive tests*
Patient characteristics			
Age	77.6	76.8	76.9
Female	0.58	0.60	0.59
White	0.88	0.9	0.90
Black	0.09	0.07	0.083
History of PE	0.003	0.006	0.02
Doctor, hospital and region characteristics			
Physician experience	16.2	15.9	16.4
	(8.4)	(8.6)	(8.6)
HRR avg spending (in $)	6,355	6,338	6,261
	(734)	(749)	(721)
Academic hospital	0.40	0.40	0.41
Top 50 research med. school	0.3	0.30	0.32
Top 50 primary med. school	0.28	0.28	0.29
No. of observations	2,380,106	78,157	5,838

Notes: Table reports means and standard deviations (in parenthesis). Data is from the Medicare claims 2000-2009, the American Hospital Association annual survey, the American Medical Association masterfile, and the Dartmouth Atlas.

Table 2: Coefficients from Testing Model and Estimated Doctor Bias

	Average	**Misweighting**	**Misweighting SE**
	(1)	(2)	(3)
Age 65-69	-0.0004	0.0026	0.0039
Age 70-74	-0.0115	0.0165*	0.0068
Age 75-79	-0.0032	0.0061	0.0041
Age 80-84	-0.0041	0.0064	0.0043
Age 85-89	-0.0068	0.0129**	0.0049
Age 90-94	-0.0166	0.0163	0.0102
Black	-0.0078	0.0230**	0.0053
Asian	0.0037	-0.0304*	0.0130
Hispanic	-0.0050	-0.0112	0.0109
Female	0.0033	-0.0069**	0.0024
Prior hospital visit w/in 30d	0.0023	0.1122**	0.0067
Prior hospital visit w/in 7d	0.0029	0.1055**	0.0091
History of pulmonary embolism	0.0005	-0.0095	0.0153
History of deep vein thrombosis	0.0088	0.0197	0.0124
Valvular disease	0.0003	-0.0386**	0.0040
Pulmonary circulation disease	0.0243	0.2756**	0.0107
Peripheral vascular disease	0.0088	-0.0206**	0.0054
Paralysis	-0.0146	0.0089	0.0104
Other nerological conditions	-0.0079	0.0444**	0.0057
Diabetes w/o chronic complications	-0.0051	-0.0082*	0.0039
Diabetes w/chronic complications	-0.0144	-0.0060	0.0102
Hypothyroidism	0.0039	0.0125**	0.0039
Liver disease	-0.0031	-0.0194	0.0128
Arthritis	0.0074	0.0002	0.0074
Coagulation deficiency	0.0024	0.0520**	0.0074
Obesity	0.0152	0.0411**	0.0076
Weight loss	0.0010	-0.0013	0.0077
Fluid & electrolyte disorders	-0.0036	-0.0063	0.0033
Blood loss anemia	-0.0167	0.0280*	0.0114
Deficiency anemias	0.0010	0.0151**	0.0040
Alcohol Abuse	0.0031	-0.0036	0.0098
Psychoses	-0.0080	0.0097	0.0100
Depression	0.0022	0.0061	0.0049
Hypertension	0.0089	0.0301	0.0040
Chronic Pulmonary Disease	0.0175	-0.0250**	0.0067
Congestive Heart Failure	-0.0003	-0.0140**	0.0036
Metastatic cancer	0.0073	0.0823**	0.0087

Notes: Column 1 reports marginal effects from coefficient estimates of the testing equation (i.e. equation 4). Recall that equation 4 is estimated separately over each of 10 deciles of physician testing probabilities; the marginal effects reported are the averages of the 10 separate estimates. Column 2 reports estimates of physicians' misweighting of these PE risk factors estimated from equation 13. Column 3 reports standard errors on these misweighting terms. *significance at the 5% level; **significance at the 1% level.

Table 3: Regressions of testing threshold on physician characteristics and practice environment

	Cross Section				Panel	
	OLS	FGLS	OLS	FGLS	OLS	FGLS
Independent variables:	(1)	(2)	(3)	(4)	(5)	(6)
Top 50 research medical school	0.0020	0.0022	0.0004	0.0008	0.0040	0.0044
	(0.0035)	(0.0032)	(0.0043)	(0.0037)	(0.0097)	(0.0079)
Top 50 primary care medical school	-0.0033	-0.0025	0.0003	0.0006	-0.0070	-0.0067
	(0.0035)	(0.0033)	(0.0043)	(0.0038)	(0.0100)	(0.0082)
Academic hospital	0.0047	0.0042	0.0056	0.0057492	-0.0064	-0.0061
	(0.0025)	(0.0023)	(0.0030)	(0.0027)	(0.0061)	(0.0053)
Log(HRR avg Medicare spending)	-0.0422**	-0.0479**	-0.0429**	-0.04802**	-0.0580*	-0.0534*
	(0.0103)	(0.0096)	(0.0125)	(0.0113)	(0.0282)	(0.0235)
Doctor experience (years)	0.0004**	0.0004**	0.0003	0.0003*	0.0005	0.0005
	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0004)	(0.0003)
For profit hospital	-0.0027	-0.0023	-0.0008	-0.0009	-0.0084	-0.0066
	(0.0034)	(0.0033)	(0.0042)	(0.0039)	(0.0088)	(0.0077)
Average income in region (in $10k)	0.0012	0.0005	0.0027	0.0015	0.0020	0.0008
	(0.0022)	(0.0020)	(0.0027)	(0.0023)	(0.0060)	(0.0048)
# of Docs	6588	6588	6588	6588	7526	7526
# of Hospitals	2695	2695	2695	2695	2695	2695
Hospital Fixed Effects	NO	NO	YES	YES	NO	NO
Doc & Year Fixed Effects	NO	NO	NO	NO	YES	YES

Notes: Each column reports results from a regression of estimated physician testing thresholds τ_d on characteristics of the physicians' training and practice environment. Even numbered columns report FGLS estimates which account for estimation error in τ_d. Columns 3 and 4 include hospital fixed effects. Columns 1 through 4 are estimated on a sample of one observed tau threshold for each doctor in sample, where an observation is a doctor. In columns 5 and 6, the observations are at the level of the physician-year, with τ allowed to vary over time for each doctor. The inclusion of this panel variation in τ allows for the estimation of physician and year fixed effects to allow identification of the effect of increasing physician experience, holding cohort fixed. All regressions control for tort reform implementation. *significance at the 5% level; **significance at the 1% level.

Table 4: Calibration Parameters

Parameter	*Value*	*Definition*	*Source*
s	0.83	test sensitivity	Lesler et al., 2009
MB_{id}	$0.025VSL$	medical benefit of testing	Lesler et al., 2009
MC_{id}	$0.0017VSL$	medical cost of testing	Lesler et al., 2009
c_{id}	$300	financial cost of testing	estimated from Medicare claims
CT	$2,800	financial cost of PE treatment	estimated from Medicare claims

Notes: Calibrated parameters of the model used in welfare simulations.

Table 5: Distribution of τ and Extent of Misweighting given VSL and False Positive Rate

		False Positive Rate			
		fp = 0	*fp = 0.03*	*fp = 0.04*	*Varies with E(fp) = .04*
		(1)	(2)	(3)	(4)
Value of a statistical life			*A. Description of Tau Distribution*		
$1 million	E(tau)	500	-241	-501	-456
	Std(tau)	812	849	862	837
	% Overtest	0.23	0.91	0.99	0.97
$2 million	E(tau)	1411	-72	-591	-502
	Std(tau)	1736	1811	1837	1782
	% Overtest	0.05	0.83	0.99	0.96
$7 million	E(tau)	5966	776	-1042	-730
	Std(tau)	6358	6619	6710	6510
	% Overtest	0.00	0.67	0.98	0.95
			B. Description of Misweighting		
$1 million	Abs(Misweight)	618	648	657	658
	SD(Misweight)	1030	1077	1094	1096
$2 million	Abs(Misweight)	1323	1379	1399	1402
	SD(Misweight)	2203	2298	2331	2334
$7 million	Abs(Misweight)	4843	5042	5111	5121
	SD(Misweight)	8068	8398	8514	8525

Notes: In Panel A, this table describes the distribution of physician testing thresholds τ_d measured in units of dollars, after correcting for estimating error using an empirical Bayes technique described in Appendix A. We report the mean and standard deviation of the τ_d distribution, as well as the fraction of physicians who are overtesting, i.e. with negative testing thresholds. Panel B reports the average absolute value and standard deviation of misweighting across patients in our sample. We report values under varying assumptions about the rate of false positive CT scan results (in columns) and the value of a statistical life (in rows).

Table 6: Patient welfare with observed testing thresholds vs. in simulations with no over-testing

	False Positive Rate					
	fp = 0		*fp = 0.3*		*fp = 0.4*	
	Actual	Sim	Actual	Sim	Actual	Sim
	(1)	(2)	(3)	(4)	(5)	(6)
% of Patients Tested	3.68%	3.52%	3.68%	2.13%	3.68%	1.91%
# of Patients Tested	57,573	55,091	57,573	33,341	57,573	29,922
Total Financial costs of testing ($ millions)	29	28	29	19	30	18
Total medical cost of testing ($ millions)	7	7	7	5	8	5
Total medical benefits of testing ($ millions)	101	100	60	53	57	49
Net benefits of testing ($ millions)	65	65	24	29	19	26
Total (financial + medical) costs per test ($)	617	629	617	713	652	763
Total benefits per test ($)	1749	1816	1035	1584	990	1632
Net benefits per test ($)	1132	1187	418	871	338	869

Notes: We compare testing behavior and social welfare under the actual distribution of physician testing thresholds τ_d (reported in odd numbered columns) to simulated behavior assuming all physicians have the optimal testing threshold of $\tau_d = 0$ (reported in even numbered columns). We report results under three different assumptions about the rate of false positive CT scan results, described in the column headers. We assume a value of a statistical life of $1 million throughout this table.

Table 7: Welfare Gains from Properly Weighting Observables

	Observed	**No Misweighting**
% of Patients Tested	3.68%	4.03%
# of Patients Tested	57573	62988
Total Financial costs of testing ($ millions)	29	37
Total medical cost of testing ($ millions)	7	11
Total medical benefits of testing ($ millions)	60	114
Net benefits of testing ($ millions)	24	66
Total (financial + medical) costs per test ($)	617	751
Total benefits per test ($)	1035	1803
Net benefits per test ($)	418	1052

Notes: We compare testing behavior and social welfare under the observed physician weighting of patient risk factors to simulated behavior assuming that physicians make no mistakes in assessing relative patient PE risk based on observable demographics and comorbidities. We assume a value of a statistical life of $1 million and a false positive rate of 0.03 throughout this table.

Table 8: Robustness of Tau Distribution to Heteroskedasticity in η

		False Positive Rate		
Model		***fp = 0***	***fp = 0.03***	***fp = 0.04***
Semiparametric w/Misweighting	E(tau)	387	-360	-622
Homoskedastic	Std(tau)	945	988	1003
	% Overtest	39	97	95
Parametric No Misweighting	E(tau)	388	-359	-620
Homoskedastic	Std(tau)	914	956	970
	% Overtest	38	97	95
Parametric No Misweighting	E(tau)	279	-449	-723
Heteroskedastic	Std(tau)	1042	1008	1097
	% Overtest	52	100	90

Notes: This table describes the distribution of physician testing thresholds τ_d measured in units of dollars, after correcting for estimating error using an empirical Bayes technique described in Appendix A. Results are reported from three different sets of modeling assumptions. First from the semiparametric model, assuming the same distribution of η_{id} across doctors and allowing for physicians to misweight observables. Second, we impose η_{id} distributed according to the uniform distribution and no longer allow for misweighting to demonstrate that the τ_d distribution is not highly sensitive to allowing for misweighting. Lastly, we allow for doctor-specific heteroskedasticity in η_{id}. For each scenario, we report the mean and standard deviation of the τ_d distribution, as well as the fraction of physicians who are overtesting, i.e. with negative testing thresholds. We report values under varying assumptions about the rate of false positive CT scan results (in columns). We assume a value of a statistical life of $1 million throughout this table.

Figure 1: Clinical Assessment of Patient with Potential Pulmonary Embolism

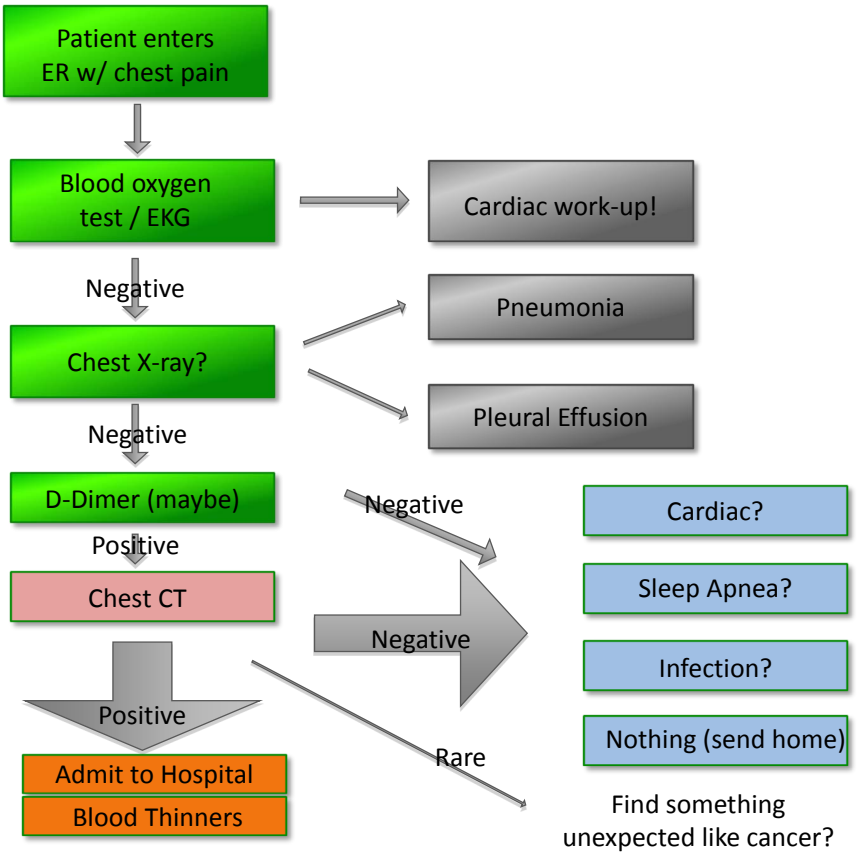


Figure 2: Net Benefits vs. Propensity to Test (fictional)

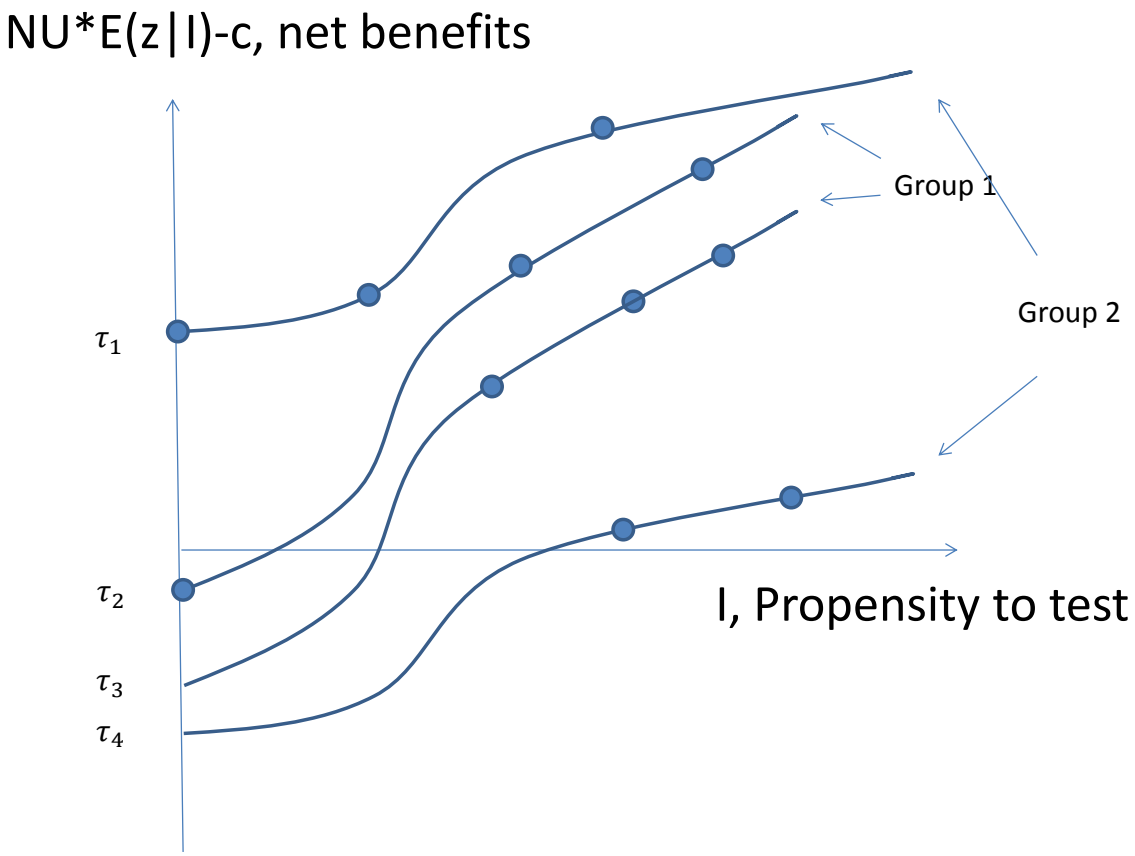
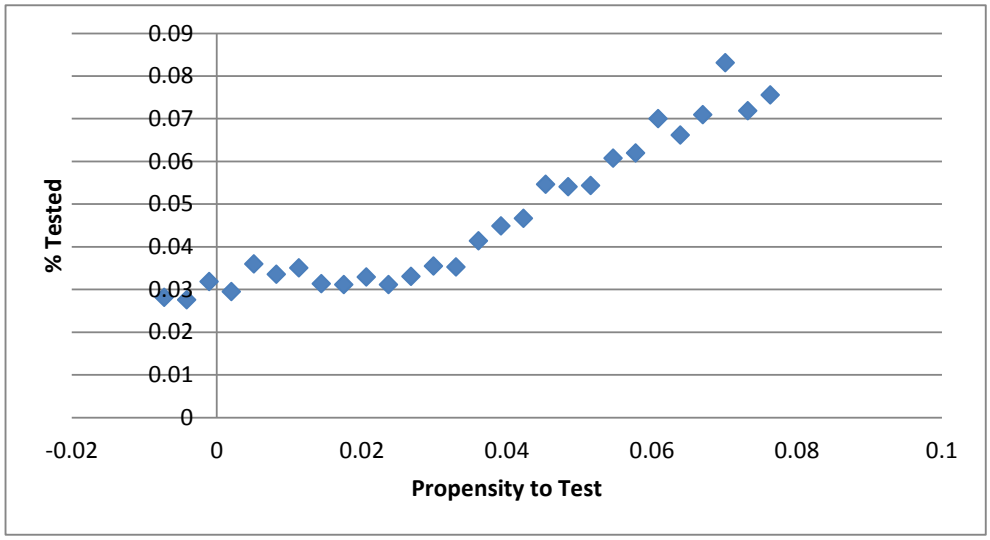
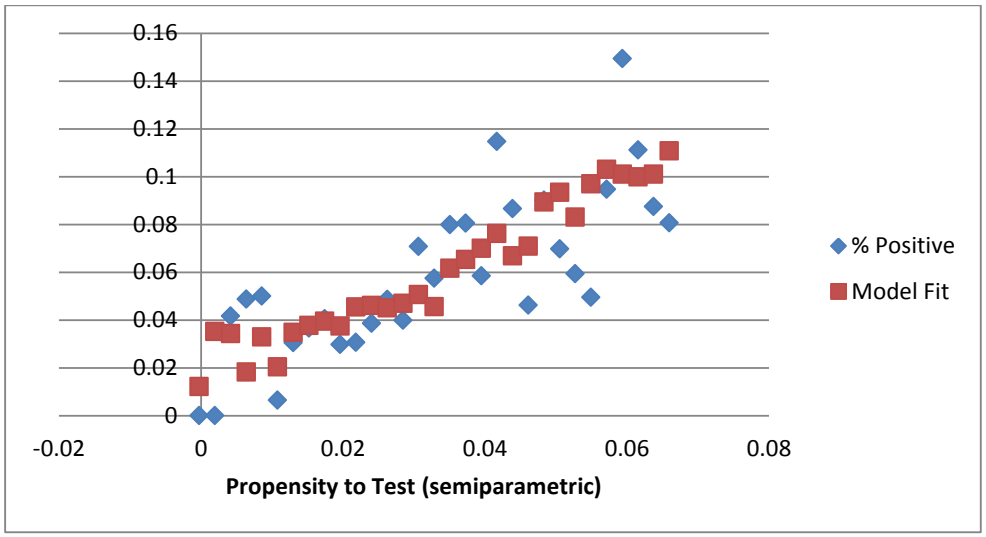


Figure 3: % of Patients who are tested vs. Propensity to Test



Notes: this figure show the relationship between the propensity to test estimated semiparametrically in decile 7 and the fraction of patients tested. The propensity is normalized so that it is zero in the 2nd percentile of all patients and normalized relative to the coefficient on one of the age dummies (in other words the location and scale of the propensity are arbitrary). The propensity is grouped into 50 cells and all cells with at least 500 patients are included.

Figure 4: % of Tests which are Positive vs. Propensity to Test



Notes: This figure plots the percentage of tests which are positive against the propensity to test among all tested patients in group 8 (of the 10 groups defined based on deciles of the proportion of patients tested). The blue points show the average % of patients tested at the propensity while the red points give the fit of the model, which consists of a common polynomial for all doctors and doctor-specific intercepts which identify the threshold parameters.

FOR ONLINE PUBLICATION

A Empirical Bayes Estimates of τ_d

In this section, we derive the estimating equation used to correct the variance of the estimated doctor fixed effects τ_d for sampling variation.

Assume that we have an equation of the form:

$$Z_{id} = \tau_d + g_q(I_{id}) + x_{id}\gamma + \epsilon_{id} \tag{24}$$

This corresponds to equation 16 (where, in the normative model, we can replace Z_{id} with $NU_{id}Z_{id} - c_{id}$). Assume that we can estimate g using polynomials of I. Then we can rewrite this equation in matrix form as:

$$Z = D\tau + X\beta + \epsilon \tag{25}$$

Let $M_x = I_n - X(X'X)^{-1}X'$ where I_n is the identity matrix. Partialling out gives:

$$M_x Z = M_x D\tau + M_x\epsilon \tag{26}$$

Let $S = M_x D$. Then our estimator of τ is given by:

$$\hat{\tau} = \tau + (S'S)^{-1}S'M_x\epsilon \tag{27}$$

For a vector x, define $var(x) = E(xx') - E(x)E(x')$ and define $var_d(x) = E(x'x) - E_d(x)^2$ (the scalar generated by taking the variance across the observations in the vector). Taking the "outer product" variance of both sides gives:

$$\begin{aligned} var(\hat{\tau}) &= var(\tau) + (S'S)^{-1}S'M_x var(\epsilon) M_x S(S'S)^{-1} && (28)\\ &= var(\tau) + (S'S)^{-1}S'var(\epsilon)S(S'S)^{-1} && (29)\end{aligned}$$

where the second line uses the fact that $M_xM_x = M_x$. Let $S^{(i)'}$ denote the ith row of S. Assuming $var(\epsilon)$ is a diagonal matrix, $S_0 = \frac{1}{N}\sum_{i=1}^{N} e_i^2 S^{(i)}S^{(i)'} \to_p \frac{1}{N}\sum_{i=1}^{N}\epsilon_i^2 S^{(i)}S^{(i)'} = \frac{1}{N}S'var(\epsilon)S$. This is asymptotically equivalent to:

$$var(\tau) = var(\hat{\tau}) - (S'S)^{-1}\left(\sum_{i=1}^{N} e_i^2 S^{(i)}S^{(i)'}\right)(S'S)^{-1} \tag{30}$$

where e_i are the residuals from equation 24. Finally, using the fact that $var_d(\tau) = \frac{1}{N_{doc}}tr(var(\tau))$ where N_{doc} is the number of doctors in our sample, we have:

$$var_d(\tau) = var_d(\hat{\tau}) - \frac{1}{N_{doc}}tr\left((S'S)^{-1}\left(\sum_{i=1}^{N} e_i^2 S^{(i)}S^{(i)'}\right)(S'S)^{-1}\right) \tag{31}$$

B Parametric Model: Uniform Distribution and Heterogeneity in Net Benefits

Let $\eta_{id} \sim U(-\eta_d, \eta_d)$. Assume further that net benefits can be written as $NU_{id} = \overline{NU} + \tilde{x}_{id}\delta$ where x_{id} are observables that we think impact the benefits of testing. As in the text, $I_{id} \equiv x_{id}\beta + \frac{\theta_{id}-c}{NU_{id}}$.

Assume the distribution of η_{id} is such that $I_{id} < \eta_d$, so there is no testing propensity I_{id} at which patients are always tested given observables (a constraint which is empirically satisfied in our data). Then, $E(\eta_{id}|\eta_{id} > -I_{id}) = \frac{\eta_d - I_{id}}{2}$. This implies that:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c}{NU_{id}} + \frac{I_{id} + \eta_d}{2} + (x_{id} - E_d(x_{id}))(\beta - \beta') \tag{32}$$

We can recover η_d up to a normalization discussed below from the testing equation. Given our distributional assumption:

$$\begin{aligned}
P(test) &= P(NU_{id}x_{id}\beta + \theta_d - c + NU_{id}\eta_{id} >= 0) \\
&= P\left(\eta_{id} >= -\left(x_{id}\beta + \frac{\theta_d - c}{NU_{id}}\right)\right) \\
&= 1 - P\left(\eta_{id} < -\left(x_{id}\beta + \frac{\theta_d - c}{NU_{id}}\right)\right) \\
&= 1 - \min\left(\frac{-(x_{id}\beta + \frac{\theta_d - c}{NU_{id}}) + \eta_d}{2\eta_d}, 1\right) \\
&= \max\left(0, \frac{1}{2} + \frac{x_{id}\beta + \frac{\theta_d - c}{\overline{NU} + \tilde{x}_{id}\delta}}{2\eta_d}\right)
\end{aligned} \tag{33}$$

In theory, η_d is identified for all doctors. In practice, for a very small number of doctors, the estimated η_d would diverge to ∞ because patients with larger $x_{id}\beta$ are less likely to be tested. These doctors are excluded from the final sample. The model with a separate η_d for each doctor is also computationally intractable, so instead we optimize for each doctor over five separate values of η_d which the model selects to maximize the likelihood function.

In the homoskedastic model reported in Section 6.1, $\eta_d = \eta_q$ is a constant within each quantile and we assume that $\delta = 0$. To estimate that model, we recover τ_d using the semiparametric procedure described in the text and use the uniform parametric assumption to recover the scaling factor between the testing and treatment equations. The propensity I_{id} in equation 35, expressed in dollars, is a linear transformation of the semiparametrically estimated propensity $\hat{I}_{id}$ in each quantile; in other words, $I_{id} = \lambda_q \hat{I}_{id} + K_q$. Recall from section 5.1 that $\hat{I}_{id}$ is normalized within each quantile q so that $\min \hat{I}_{id} = 0$ among tested patients. In the model, the minimum value of I_{id} consistent with testing is $-\eta_q$. Thus, $-\eta_q = K_q$. Plugging into equation 35 gives:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c}{NU_{id}} + \frac{\lambda_q \hat{I}_{id}}{2} + (x_{id} - E_d(x_{id}))(\beta - \beta') \tag{34}$$

Therefore, the scaling factor λ_q in each quantile can be recovered from a regression of the function

$h_q(I_{id}) = NU_{id}(I_{id} + g_q(I_{id})) = \frac{\lambda_q \hat{I}_{id}}{2}$ on $\hat{I}_{id}$, which is the approach taken in the text.

In the heteroskedastic model, I_{id} and all the η_d are identified only up to a normalization, but the same approach can be used to estimate the scaling factor. Let $\hat{I}_{id} = x_{id}\hat{\beta} + \frac{\hat{\theta}_d - c}{NU_{id}}$ and $\hat{\eta}_d$ denote the estimated η (given our normalization of $\beta_1 = 1$). Then we have: $I_{id} = \lambda \hat{I}_{id}$ and $\eta_d = \lambda \hat{\eta}_d$ (the scaling constant is the same for each since the ratio of I_{id} and η_d is identified). This implies:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c}{\overline{NU} + \tilde{x}_{id}\delta} + \frac{\lambda(\hat{I}_{id} + \hat{\eta}_d)}{2} + (x_{id} - E_d(x_{id}))(\beta - \beta') \tag{35}$$

where now τ_d can be estimated simultaneously with the scaling factor λ in a constrained least squares regression.

Intuitively, heteroskedasticity in η_d is identified by the fact that observables are less predictive of testing behavior for doctors with more private information

Using this model, we can compute the welfare cost of overtesting as follows. Let $\hat{t}_{id}(\tau_d, \Delta\beta)$ denote the probability that consumer i is tested by doctor d as a function of τ_d and $\Delta\beta = \beta - \beta'$ and let $\hat{Z}_{id}(\tau_d, \Delta\beta)$ denote the probability of a positive test conditional on testing. To compute $\hat{t}_{id}(0, \Delta\beta)$, we use the fact that $I(0, \Delta\beta) = I(\tau_d, \Delta\beta) + \frac{\tau_d}{NU_{id}}$ so $\hat{I}(0, \Delta\beta) = \hat{I}(\tau_d, \Delta\beta) + \frac{\tau_d}{\lambda NU_{id}}$. Then we have: $\hat{Z}_{id}(0, \Delta\beta) = \frac{c}{NU_{id}} + \frac{\lambda_q \hat{I}_{id}(0, \Delta\beta)}{2} + (x_{id} - E_d(x_{id}))(\beta - \beta')$. Likewise, to compute $\hat{t}_{id}(\tau_d, 0)$ (the propensity to test with no misweighting), we use the fact that $I(\tau_d, 0) = I(\tau_d, \Delta\beta) + (x_{id} - E_d(x_{id}))\Delta\beta$ so $\hat{I}(\tau_d, 0) = \hat{I}(\tau_d, \Delta\beta) + \frac{(x_{id} - E_d(x_{id}))\Delta\beta}{\lambda}$. Then we have: $\hat{Z}_{id}(\tau_d, 0) = \frac{\tau_d + c}{NU_{id}} + \frac{\lambda_q \hat{I}_{id}(\tau_d, 0)}{2}$.

We define the following objects:

$$\begin{aligned}
MB(\tau_d, \Delta\beta) &= \sum_i P(test_i) \cdot P(PE_{id}|test) MB_{id} = \sum_i \hat{t}_{id}(\tau_d, \Delta\beta) \frac{s(\hat{Z}_{id}(\tau_d, \Delta\beta) - fp)}{(s - fp)} MB_{id} \\
MC(\tau_d, \Delta\beta) &= \sum_i P(test_i) P(Z_{id} = 1|test_i) MC_{id} = \sum_i \hat{t}_{id}(\tau_d, \Delta\beta) \hat{Z}_{id}(\tau_d, \Delta\beta) MC_{id} \\
FC(\tau_d, \Delta\beta) &= \sum_i P(test_i)(c + P(Z_{id} = 1|test_i) CT_{id}) = \sum_i \hat{t}_{id}(\tau_d, \Delta\beta)(c + \hat{Z}_{id}(\tau_d, \Delta\beta) CT_{id}) \\
NB(\tau_d, \Delta\beta) &= MB(\tau_d, \Delta\beta) - MC(\tau_d, \Delta\beta) - FC(\tau_d, \Delta\beta)
\end{aligned} \tag{36}$$

where MB denote the medical benefits of testing (derived in section 5.3.1, MC denotes the medical costs of testing, FC denotes the financial costs of testing and NB denotes the net benefits of testing as a function of these objects. We define the welfare cost of overtesting as $NB(0, \Delta\beta) - NB(\hat{\tau}_d, \Delta\beta)$ and the welfare cost from misweighting as $NB(\tau_d, 0) - NB(\hat{\tau}_d, \Delta\beta)$ where $\hat{\tau}_d$ is drawn from the estimated empirical Bayes distribution.

C Testing for Multiple Conditions

An important caveat to our above analysis is that claims data is only sufficient to identify CPT codes for "chest CT with contrast"; we cannot isolate CT scans that follow the PE testing protocol specifically. Although tests for PE are the primary indication for chest CTs in the emergency room setting, there are other possibilities. Because of this limitation, some of the tests we have labeled as "negative" since the patient is not diagnosed with pulmonary embolism may be tests

performed for a different indication. There are five main alternative indications for CT scans in an emergency department setting: trauma, lung or chest cancers, aortic dissection, pleural effusion, and pneumonia. We discuss our approach to each of these alternative diagnoses in turn.

We exclude from the estimation sample patients with diagnosis codes related to trauma (such as fractures, injury, motor vehicle accidents), when these codes are associated with bills on the same day as the patient's emergency department evaluation. Chest CTs for these patients are likely aiming to assess damage from a trauma rather than a pulmonary embolism. In a detailed sample of patient records from chest CT scans performed in the emergency room of a large hospital, diagnosis codes associated with the radiology bills readily distinguished traumas from other scanning indication.

Similarly, we exclude patients with a history of aortic aneurysm, aortic dissection, or other arterial dissection, in order to eliminate patients for whom chest CTs may be intended to evaluate for aortic dissection. Aortic dissections are extremely rare, with only approximately 9000 cases per year in the United States, making it over 30 times less common than pulmonary embolism (Meszaros et al. 2000).

It is unusual for a cancer diagnosis to be made for the first time in the ED, but patients with worsening symptoms as a result of tumor growth or metastasis and occasional new diagnoses may be seen. CT scanning is routinely used to diagnose and stage cancers. In our sample of detailed ED chest CT records from the academic medical centerl, fewer than 1% of the scans were used to diagnose or stage cancers. In the Medicare data, we exclude those patients with chest cancer indicated on their visit to the emergency room or associated inpatient visit from our preferred estimation sample.

Chest CTs can be used to guide a procedure to treat patients with pleural effusion, which is typically first diagnosed with a chest X-ray. Because a chest CT is not commonly a diagnostic test for pleural effusion but rather an input into the treatment of the disease, we can exclude patients from the sample with diagnoses of pleural effusion. Since some patients are diagnosed with both pleural effusion and pulmonary embolism, and in these patients the chest CT was likely serving a diagnostic role, we do not exclude pleural effusion patients with a diagnosis of pulmonary embolism. These sample restrictions will tend to overstate the rate of positive testing and bias us away from finding evidence of over-testing, since we may be excluding some pleural effusion patients who are being tested for pulmonary embolism but have a negative test result.

Together, these exclusions for patients with trauma, cancer, or pleural effusion remove 32% of patients receiving chest CTs from our sample. Results presented above are qualitatively similar when these patients are included.

Finally, chest CTs can be used to diagnose pneumonia. Pneumonia can also be reliably diagnosed with cheaper and lower radiation technologies (David et al. 2012); the added value of a chest CT with contrast in an ED setting for diagnosing these alternative conditions is very modest (Venkatesh et al. 2013). Technically, the value of a chest CT scan for diagnosing a condition that could otherwise be detected with an X-ray is bounded by the costs of the x-ray, which is about $30 in our sample. Accounting for a $30 additional net benefit from diagnosing pneumonia whern indicated does not substantively change our results about the welfare costs of overtesting.