Negative Tests and the Efficiency of Medical Care: Investigating the Determinants of Imaging Overuse

Jason Abaluck and Leila Agha

with Chris Kabrhel, Ali Raja, Jae Lee, Arjun Venkatesh

*

January 10, 2013

Abstract

There is enormous variation in medical treatment across physicians, hospitals and regions but designing reforms to lower costs and maintain quality requires identifying specific instances of inefficient spending. We develop a measure of the efficiency of health care delivery based on the frequency of negative CT scans for pulmonary embolism. Our model shows how to transform the fraction of negative tests into a measure of medical care efficiency that links directly to welfare. We apply our model using a 20% sample of Medicare claims data from 2000-2009; the empirical assignment of testing outcomes is validated using chart and billing data from two large hospitals. We find that 80% of doctors are performing too many tests, in the sense that on the margin they perform tests even if the costs exceed the benefits. If all doctors tested only when the benefits exceeded the costs, the proportion of patients given a chest CT in our sample would fall by 15%, from 3.63% to 3.08%. The financial savings would be about $66 per person tested, while the medical benefits due to reduced mortality risk from treatment of false positives would be $242 per person tested; together, these factors would roughly double the welfare increase from testing over a world with no treatment. We also find that more experienced doctors and doctors in regions with lower spending overall are less likely to overtest. We use the across region relationship between total spending and wasteful CT spending controlling for non-wasteful CT spending to estimate the overall prevalence of wasteful spending in health.

*Thanks to Brian Abaluck, Joe Altonji, Joshua Aronson, Jonathan Gruber, Nathan Hendren, Mitch Hoffman, Lisa Kahn, Danielle Li, David Molitor, Constana Esteves-Sorenson, Ashley Swanson, and Blair Parry. Funding for this work was provided by NIA Grant Number T32-AG0000186 to the NBER.

1 Introduction

There is enormous variation in medical treatment across physicians, hospitals and regions but designing reforms to lower costs and maintain quality of care requires identifying specific instances of inefficient spending. 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). And yet, there is a growing body of evidence that higher spending, resource intensive hospitals and regions do achieve better health outcomes at least in the context of high acuity, emergency care (Doyle 2007 and Doyle et al. 2012). Untargeted cuts may lead to worse outcomes, underscoring the importance of identifying specific instances of wasteful spending.

In this paper, we develop a measure of the efficiency of medical resource utilization based on the frequency of negative CT scans for pulmonary embolism. A doctor who performs many negative CT scans, which have little *ex post* value for improving patient health, is likely over-using this test. The optimal fraction of negative tests may vary across doctors depending on the ex ante propensity of the patient population to develop a given condition and the benefits of treatment if a test is positive. Given this patient heterogeneity, our model shows how to transform the fraction of negative tests into a measure of medical care efficiency that links directly to welfare. Our model can be estimated using only claims data (as opposed to more detailed chart data) which makes it possible to study overuse in the full population of medicare beneficiaries. This large sample size allows us to investigate the determinants of overuse: we study how medical training, reimbursement, malpractice law, hospital characteristics (for-profit and teaching hospitals) and regional characteristics such as spending impact the efficiency of medical care delivery.

We build on the theoretical framework of Chandra and Staiger (2011) (hereafter CS) who use a structural model to estimate the medical returns to heart attack treatment and decompose variation in utilization into differences in physician skill, patient population, and propensity to over-use medical intervention. CS assume that doctors treat if the net benefits exceed a doctor-specific threshold τ_d and in structural estimation they seek to recover τ_d. A value of $\tau_d < 0$ would indicate that a doctor is willing to treat even if the net benefits of doing so are negative; i.e. it would indicate overtreatment.

In most contexts, measuring overuse requires estimates of the effect of treatment on the treated for each patient; ideally, this parameter would be estimated using randomized variation or credible

instruments. CS argue that reliable estimates 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.

A key insight of this paper is the adaptation of this model to the context of medical testing, where in the case of chest CT scans, the *ex post* value of the test to the patient is partially observable in insurance claims records based on whether the test leads to the diagnosis being tested for. This innovation allows us to develop a doctor-specific measure of care overuse in a large, national sample of physicians and patients. We can then examine the correlates and determinants of care overuse, by estimating the relationship between a physician's utilization threshold and his training or practice environment.

We apply our model using a 20% sample of Medicare claims data from 2000-2009 which we validate by comparing billing data with patient records at two large hospitals in Boston. Given our measure of inefficient testing, we investigate many questions about the determinants of effective medical care. In reduced form regressions of the indicator for negative tests on doctor and regional characteristics, we find that less experienced doctors and higher spending regions are more likely to order negative tests.

In our structural model, we find that nearly 80% of doctors are overtesting in the sense that for their marginal patient, the costs of testing exceed the benefits. We find that the reduced form results are reflected in the structural model: less experienced doctors and higher spending regions are more likely to overtest. We also use the model to conduct several welfare analyses. If all doctors behaved optimally, the total benefits to patients from chest CTs would roughly double and spending on CT scans and patient admissions would fall by 12%.

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 5 lays out our structural model of testing behavior and derives an equation relating the indicator for positive and negative tests to the threshold τ_d described above which indicates whether or not a doctor is an overtester. Section 4 reports the reduced form results from a regression of the indicator for a positive test on covariates (a regression motivated by the structural model) and section 6 describes how we estimate the model and reports some results. Section 7 examines the robustness of the structural model, section 8 conducts the various welfare exercises

described and section 9 concludes.

2 Context

We study testing behavior in the context of chest CT scans performed in the emergency room (ER) 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. 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 mild PE (Lessler, Isserman, Agarwal, Palevsky, and Pines 2010).

This test has a number of attractive features for our purposes: it is a frequently performed test; it introduces 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: an alternative diagnosis is less likely than PE, the patient has an elevated heart rate, patient was immobilized for at least three days or underwent surgery in the previous month, or the patient has a history of deep vein thrombosis or pulmonary embolism.

Despite these guidelines, many argue that PE CT scans are widely overused (Coco and O'Gurek 2012, Mamlouk, vanSonnenberg, Gosalia, Drachman, Gridley, Zamora, Casola, and Ornstein 2010 and Costantino et al. 2008). The American College of Radiology targeted PE CT in one of its five recommendations for reducing the misuse of imaging, as part of the Choosing Wisely campaign aimed to reduce overuse of medical services (American College of Radiology 2012). The nonspecific symptoms and significant mortality risk likely both contribute to the overuse, particularly in the ER 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 rolled into a capitation payment. In addition to this financial cost, testing comes 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 carries its own financial costs and significant risk of internal bleeding.

2.2 Identifying PE CTs in claims data

We can identify testing for a PE in the Medicare claims data, using bills submitted by radiologists for the interpretation of chest CTs with contrast on the same day as an ER visit. 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, pneumonia, and traumas.

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 indicated on either their Medicare Part B bills submitted the same day as the emergency room evaluation or any ensuing inpatient stay bill. 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. We explore robustness of the model to including pleural effusion diagnoses or considering chest CT as a diagnostic test for pleural effusion in section 7.

We also exclude from the sample patients with diagnosis codes related to trauma (such as fractures, injury, motor vehicle accidents) and codes related to chest or lung cancers, when these codes are associated with bills on the same day as the patient's emergency room evaluation. Chest CTs for these patients are likely aiming to assess damage from a trauma or growth of a cancer rather than a pulmonary embolism. Results presented below are qualitatively similar when these patients

are included. Together, these exclusions for patients with trauma, cancer, or pleural effusion exclude 32% of patients receiving chest CTs in this sample.

Finally, chest CTs can be used to diagnose pneumonia. In the absence of any clinical suspicion for an alternative diagnosis such as pleural effusion or pulmonary embolism, pneumonia can be accurately diagnosed with less medical risk and financial cost with an x-ray. If a physician was concerned about both pneumonia and pulmonary embolism, then it is possible that he would substitute to a chest CT rather than an x-ray. Assuming that either a chest x-ray or chest CT would always be performed when in the case of clinical suspicion of pneumonia, the value of the chest CT derives solely from the increased probability of diagnosing a pulmonary embolism when the CT is utilized and the cost of the test is simply the additional financial and medical costs of performing a CT, over and above the costs of performing an x-ray. Since the health costs and financial costs of an x-ray are much, much lower than the costs of a CT scan (machinery is comparatively inexpensive, faster to interpret, radiation dose is much lower, and there is no risk of a contrast reaction), this cost adjustment is minor. In light of this context, we include pneumonia patients in the baseline sample and do not code pneumonia as a positive test.

2.3 Identifying positive CT scans

In addition to identifying CT scans in billing data, we also 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. Thus, we identify positive tests on the basis of Medicare Part A hospital claims that include diagnoses of pulmonary embolism.

We have validated this model of identifying positive tests by using cross-referenced patient chart and hospital billing data from a large tertiary care hospital (hereafter LTC hospital). In particular, we may undercount 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.

In LTC hospital sample, we found that 90% of patients who test positive for PE in the emergency room were admitted within 1 day. Patients with very mild PE's may occasionally be discharged 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 ER do not have an inpatient claim with a PE diagnosis.

Of patients with a negative PE CT scan recorded in their emergency room 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 ER workup. This error could occur if the patient develops a PE later in their 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 ER PE CT. Of the 94% of tests we code as negative, 1.1% of the patients had positive ER 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 Data

3.1 Medicare Claims

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 room on the basis of physician submitted bills. Using physician identifiers, we track the behavior of all doctors who routinely evaluate Medicare

patients in the emergency room.

In the sample of Medicare patients evaluated in the emergency room, we measure whether each patient was tested with a chest CT scan within one day of their emergency room evaluation using Medicare Part B bills for following CPT codes: 71260, 71270, and 71275. This is the primary measure of testing used in our analysis. We indicate a patient as having a positive test if they are admitted to the hospital with a diagnosis of pulmonary embolism indicated as a primary or secondary diagnosis code on the Medicare Part A bill for their 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. In addition to including a standard set of 30 medical comorbidities (following Elixhauser et al. 1998), we include several measures that are specific to PE risk.[1] 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.

3.2 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 over-testing 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.

[1] Conditions are defined using a 1-year inpatient medical history, based on Medicare Part A institutional claims. These diagnosis 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.

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.

Data on state tort reform is from Avraham (2011) Database of State Tort Law Reforms. We use this database to measure whether a state has enacted malpractice damage caps on award amounts, or joint and several liability reform. These data allow a difference-in-differences style analysis of how practice patterns change when state's limit physician's exposure to malpractice risk.

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.

3.3 Summary Statistics

There are over 6 million emergency room visit evaluations in our dataset, after excluding patients with trauma, chest cancer, and pleural effusion diagnoses. Of these patients evaluated in the ER for any reason, 2.2% of them are tested with a chest CT scan with contrast. Amongst tested patients, 6.4% 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 65,000 physicians, with an average of over 90 ER 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 7 times more likely to have a history of pulmonary embolism.

Patients with negative tests are evaluated by doctors with 7 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 1.5% higher amongst negative tested

patients. 36% 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.

4 Reduced Form Estimation

4.1 Reduced form model

Reduced form regressions estimate the relationship between a doctor's testing outcomes and his training, experience, and practice environment. The idea is that a doctor who orders many tests that turn out to be negative likely has a low threshold for when it is worthwhile to test. The regressions are estimated over the sample of tested patients, and they take the following form:

$$Z_{id} = a_1 + a_2 Y_d + a_3 X_i + \epsilon_i d \tag{1}$$

Z_{id} is the testing outcome for patient i evaluated by doctor d; it equals one if the patient is diagnosed with a pulmonary embolism. Y_d is a vector of doctor characteristics, including his experience and training and the type of hospital and region that he practices in. X_i is a vector of patient characteristics. Standard errors are clustered by hospital referral region.

Some variability in testing outcomes may be introduced by differences in the patient characteristics; for example, a doctor who sees more patients with a history of deep vein thrombosis is likely to both test more and receive more positive test results. Each patient has an array of characteristics, some observable by the econometrician, others observable only to the doctor, that contributes to his risk for pulmonary embolism. If conditional on the observables controlled for in the regression, each doctor faces the same distribution of patients, then we can attribute conditional differences in testing behavior to the physician's testing propensity to differences in the testing threshold. We show that the results presented here are similar across specifications that vary in how richly patient characteristics are controlled for, providing some evidence that conditional heterogeneity in the patient population is not driving the observed differences in testing outcomes.

There are a three main limitations of the reduced form approach. First, if there remain unobservable differences across doctors in their patients' ex ante risk for pulmonary embolism, then we may mistakenly attribute differences in the patient's risk profile to the doctor's testing threshold. The structural model addresses this concern allowing different doctors to treat patient populations

that unobservably differ, on average, in their ex ante risk. In refinements to the basic structural model, we further relax this assumption by allowing doctors to face not just different unobservable patient characteristics on average but also heteroskedastic distributions of unobservable patient risk.

Second, the above regression assumes that the benefits of treatment do not vary across doctors and patients, so that there would be no reason for doctors to differ in their testing behavior once we've conditioned on the patient's risk of having a PE. In the structural model, we allow for the fact that patient characteristics may impact the benefits from treating a PE.

Lastly, while the reduced form model allows us to estimate differences in testing thresholds under the assumptions outlined above, it does not allow welfare analysis. We cannot distinguish doctors that are over- or under-testing, nor make any normative statements about whether changes to the testing rates would be welfare enhancing. Using calibrated assumptions about the value of testing and treatment drawn from the medical literature, the structural model does allow for the identification of over-testers.

4.2 Reduced form results

Regression results are reported in Table 5. Results are highly consistent across all specifications (with one exception, noted below); we focus on column 4 in describing the results, since this includes data from all physicians and years, along with the richest set of controls including state fixed effects, patient comorbidities and pulmonary embolism risk factors.

Doctor experience, defined as the number of years since graduating medical school, is strongly correlated with the probability that the patient has a positive CT scan. The finding suggests that every 10 additional practice years of the ordering physician is associated with a 0.43 percentage point increase in a positive CT finding, from a mean of 6.4 percent positive tests, significant at the 1% level.

The experience profile is further unpacked in Figure 1, where we see that doctors with 0-4 years of experience are most likely to order a CT scan that turns out to be negative for pulmonary embolism. Rates of positive tests steadily improve until the doctor has 20-29 years of experience, which is statistically indistinguishable from 30-39 years of experience. Very old doctors with 40 or more years experience begin ordering more negative tests again, although due to the small sample of physicians still practicing at that age, the estimate is imprecise.

Note that due to the high degree of correlation between age and experience (or alternatively,

cohort and experience), we do not have sufficient power to statistically distinguish these mechanisms in the data. We observe fewer than five CT scans per in-sample doctor, on average, so despite the panel structure of the data, we cannot estimate a precise experience profile after controlling for physician fixed effects. However, the strong correlation between experience (or age) and testing behavior is suggestive of two possible mechanisms: a strong learning effect, where doctors raise their testing threshold over time or learn to distinguish more finely between low- and high-risk patients; or, a notably different practice style by physician cohort, where older physicians or those born in earlier years are less inclined to pursue testing for low-risk patients.

Average medical spending within the HRR is also strongly related to the probability that a tested patient has a pulmonary embolism. This data is merged from the Dartmouth Atlas, and gives the average spending per Medicare beneficiary, adjusted for age, race, sex, and price, from 2003–2009. In column 4, a ten percent increase in regional spending levels is associated with 0.43 percentage point decline in the probability of a positive test amongst tested patients, significant at the 1% level. This finding provides suggestive evidence that some of the raw variation in Medicare spending across regions may be driven by differences in wasteful spending.

There is also evidence that in difference-in-differences estimation, controlling for time and state fixed effects, the enactment of joint and several liability reform in a state is associated with an increase in the rate of positive tests for PE. This evidence supports a defensive medicine mechanism, whereby the threat of a lawsuit makes physicians more likely to order a test on patients with low expected benefits. We are currently working on estimating the structural model to identify panel variation in physician or hospital's testing behavior over time. This will allow us to extend the difference-in-differences identification strategy to the analysis of the structural model results.

Lastly, we find no significant impact of whether the physician typically practices at an academic hospital, physician gender, or physician's medical school quality. These coefficients are imprecisely estimated and not statistically distinguishable from zero in any of the reduced form specifications.

5 Model of Testing Behavior

We will now develop a model of physician's testing decisions and test outcomes that allows us to identify whether doctors are under- or overusing medical tests. A doctor must decide whether or not to test each patient he evaluates in the emergency room 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).

The starting point for our model is the assumption that doctors test a patient only if the perceived net benefits of testing given all of the information available to them 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 who 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 doctor threshold parameters, τ_d, can be recovered without separately estimating the net benefits of treatment for each patient. It is sufficient to know whether the test was positive or negative even if the net benefits are allowed to vary flexibly based on patient's medical histories.

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.

To motivate the full model, consider first a simplified case where the net benefits of testing are equal to the probability of a positive test, q_{id}, and assume there is no heterogeneity in the benefits of treatment across patients who tested positive. The probability of a positive test may vary with observable patient characteristics: $q_{id} = x_{id}\beta + \eta_{id}$, where 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. For now, we assume that

η_{id} is i.i.d. across doctors and patients.[2] Further, assume that the costs of testing are a known constant c. Doctors will test if $q_{id} - c \geq \tau_d$.

Under these assumptions, we could estimate the equation $P(test) = f(x_{id}\beta - \tau_d - c)$ and immediately recover τ_d. In other words, if doctor A tests more patients than doctor B, conditional on observable patient characteristics, then we can immediately infer that doctor A has a lower testing threshold. We would not need to observe testing outcomes to recover τ_d.

This basic model assumes that there are no unobservable differences in patient mix across physicians. This is a strong assumption in the medical context where the rich variation in patient risk across doctors is thought to be difficult to observe from claims data. 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 level 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.

In this model, $P(test) = f(x_{id}\beta + \alpha_d - \tau_d - c)$, so the testing equation is only sufficient to identify $\theta_d = \alpha_d - \tau_d$: we cannot tell if a given doctor tests a lot given observables because she is an overtester (small τ_d) or because he has a patient population which is particularly predisposed to pulmonary embolysm (large α_d). All is not lost however because we can distinguish α_d and τ_d if we also observe the number of *positive* tests for a given doctor. If a doctor tests more patients given observables because they have a large α_d (as opposed to a small τ_d), then they should also produce more positive tests.

That is the fundamental intuition of our model. Let us now lay out a more complete version. Given the assumption that negative tests are not *ex post* medically valuable, the net benefits of testing are given by the the doctor's perceived probability of a positive test (q_{id}) times the net utility conditional on treatment NU_{id} minus the cost of testing, c_{id}. Together, these assumptions

[2]This is all we do in this draft. We can in principle estimate a model with heteroskedastic η provided we make a parametric assumption about η_{id} (e.g. that it is normally distributed)—this allows some doctors to do a better job of deciding which patients to test given observable x's as opposed to just having different thresholds.

imply that doctor d tests patient i if and only if:

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

We assume that net utility of treatment, given the patient has tested positive, is given by:

$$NU_{id} = \overline{NU}_{id} + \tilde{x}_{id}\delta \tag{4}$$

where $\overline{NU}_{id}$ is a known component of net utility which we compute directly for each patient based on their medical history and $\tilde{x}_{id}$ includes observables which may impact net utility but whose relationship to net utility is estimated in the model.

Define $\theta_d = \overline{NU}_{id}\alpha_d - \tau_d$. Plugging our specifications for the probability of a positive test and for net utility into the testing equation and rearranging yields the final form of the testing equation:

$$x_{id}\beta + \frac{\theta_d + \alpha_d\tilde{x}_{id}\delta - \tilde{c}_{id}}{\overline{NU}_{id} + \tilde{x}_{id}\delta} + \eta_{id} \geq 0 \tag{5}$$

This yields a standard semiparametric binary choice model of testing. We next show how the 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 $I_{id} \equiv x_{id}\beta + \frac{\theta_d + \alpha_d x_{id}\delta - \tilde{c}_{id}}{\overline{NU}_{id} + \tilde{x}_{id}\delta}$. From equation 5, we can compute the expected benefits conditional on testing, which are given by:

$$E(B_{id}|T_{id} = 1) = \tau_d + (\overline{NU} + \tilde{x}_{id}\delta)I_{id} + (\overline{NU} + \tilde{x}_{id}\delta)g(I_{id}) \tag{6}$$

where $g(I_{id}) = E(\eta_{id}| - \eta_{id} \leq I_{id})$ is an (unknown) function of I_{id}.

Let Z_{id} be an indicator for whether a test was positive or negative. If doctors have rational expectations, we must have $E(q_{id}|T_{id} = 1) = E(Z_{id}|T_{id} = 1)$. Given these rational expectations and equation 3, we can write the expected benefits as $E(B_{id}|T_{id} = 1) = (\overline{NU} + \tilde{x}_{id}\delta)E(Z_{id}|T_{id} = 1) - c_{id}$. Plugging this into equation 6 and rearranging yields:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c_{id}}{\overline{NU} + \tilde{x}_{id}\delta} + I_{id} + g(I_{id}) \tag{7}$$

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}.

Intuitively, imagine that doctor A and doctor B have observably similar patients and the same propensity to test. This might be because they have comparable tendencies to overtest, i.e. the same testing threshold τ_d, in which case they must also have comparable α_d's; in this case, we would observe the same rates of positive test results amongst tested patients of both doctor A and doctor B. Alternatively, it might be that doctor A overtests relative to doctor B (τ_A¡τ_B) but doctor A also has a less suitable patient population (α_A¡α_b), so that on net he has the same propensity to test as doctor B. In this later case, we should observe fewer positive tests for doctor A. Thus, in equation 7, we see that controlling for the propensity to test, a doctor with fewer positive tests will have a smaller τ_d since only a (relative) overtester would have the same propensity to test for a less suitable patient population.

Note additionally that it is sufficient to observe only whether the test is positive or negative and not the patient-specific benefits of treatment because heterogeneity in the net utility of treatment conditional on a positive test can also be recovered from the testing equation. This is because the impact of the net utility of treatment on the testing decision should scale with the likelihood of a positive test. If we observe that doctors are differentially inclined to test patients with large x's when other observable factors make a positive test more likely (controlling for the direct impact of x on the frequency of positive tests), this suggests that net utility varies with x.[3]

As noted above, the estimation of equation 7 only allows the identification of testing thresholds τ_d up to a constant normalization. From this equation, we cannot directly recover the absolute magnitudes of τ_d—we can say whether doctor A appears to have a higher threshold than doctor B for deciding which patients to test, but we cannot say whether both doctors are testing too little (meaning doctor B is doing relatively better), whether doctor A is testing too little and doctor B testing too much, or whether both doctors are testing too much (meaning doctor A is doing relatively better). In section 6.2, we discuss how we can determine the appropriate normalization for the estimated τ_d and thus determine which physicians are over-testing and which are under-testing.

Equation 7 motivates the reduced form exercise in the previous section of regressing the indicator

[3]A somewhat subtle point is that only because of heterogeneity in τ_d can we separately identify heterogeneity in net utility (which scales with α_d) from non-linearities in the function relaxing the propensity to consume to the probability of testing. Thus, δ is only separately identified when equations 6 and 7 are estimated jointly.

for a positive test on doctor, hospital and regional variables controlling for patient characteristics. In particular, if we controlled in a sufficiently flexible way for all patients, if there were no doctor level unobservable variables which impacted the probability of a positive test ($\alpha_d = 0$), and if any variation in net utility across patients were completely observable and did not need to be estimated ($\delta = 0$), then the reduced form exercise would recover exactly the same parameters as a regression of the τ_d thresholds on covariates scaled by $\overline{NU}$.

6 Calibration and Estimation of the Structural Model

6.1 Calibration of Parameters

To estimate the model laid out in Section 5, we need to determine the values of c_{id} 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, Fowler, Goodman, Gottschalk, Hales, Hull, Kenneth V. Leeper, John Popovich, Quinn, Sos, Sostman, Tapson, Wakefield, Weg, and Woodard 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_{id} denotes the medical costs of treatment and CT_{id} denotes the financial cost of treatment. Then the (known) component of the net utility of a positive test is given by:

$$\overline{NU}_{id} = P(PE_{id}|Z_{id} = 1)(MB_{id} - MC_{id}) + (1 - P(PE_{id}|Z_{id} = 1))(-MC_{id}) - CT_{id} \tag{8}$$

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

$$\overline{NU}_{id} = \frac{s(q_{id} - fp)}{q_{id}(s - fp)}(MB_{id} - MC_{id}) + (1 - \frac{s(q_{id} - fp)}{q_{id}(s - fp)})(-MC_{id}) - CT_{id}) \tag{9}$$

We can therefore write the net benefits of testing as:

$$\begin{aligned} B_{id} &= q_{id}(\overline{NU} + x_{id}\delta) - c_{id} \\ &= \frac{s(q_{id} - fp)}{(s - fp)}(MB_{id} - MC_{id}) + (q_{id} - \frac{s(q_{id} - fp)}{(s - fp)})(-MC_{id}) - CT_{id}q_{id} + q_{id}x_{id}\delta - c_{id} \end{aligned} \quad (10)$$

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

$$B_{id} = q_{id}(\hat{NU}_{id} + x_{id}\delta) - \hat{c}_{id} \quad (11)$$

which is exactly the definition of net benefits in Section 5. Conditional on whether or not testing and treatment are observed, false positives and false negatives impact only marginal benefits; that is, the costs of testing are paid if a test is done and the costs of treatment are paid if treatment is performed, but the marginal benefits of treatment accrue only if the patient actually has the underlying condition. 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 3. 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_{id} will be proportionately less important (and so testing will be more desirable). In our baseline estimates, we use a VSL computed as a function of life-expectancy given age where remaining years are valued at \$100,000 per life year. This yields an average in our sample of about \$1 million per patient. (In forthcoming results, we show that our main results are not altered by using values at the lower or upper end of VSL estimates–\$1 million and \$7 million respectively.)

6.2 Who is an overtester?

As noted in Section 5, equation 6 only allows us to recover the relative values of τ_d—we also need to determine an appropriate normalization in order to identify which doctors are overtesters and which are undertesters. In other words, let τ_d^* denote the true τ's and $\hat{\tau}_d$ the τ's estimated from the model. We know that $\tau_d^* = K + \hat{\tau}_d$ and we want to determine the constant K.

To do so, we examine expected benefits for "marginal patients". The expected benefits for the average patient will be greater than the threshold value for τ_d since doctors test if and only if $B_{id} \geq \tau_d$. But by computing expected benefits for patients whose doctors are just indifferent

between testing and not testing, we can recover τ_d (which by definition is equal to the expected net benefits for the marginal patient). 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 cannot be tested. Further, at $\underline{I}$, we know that $\eta_{id} = \overline{\eta}$. In other words,

$$\lim_{I \to \underline{I}} g(I_{id}) = \lim_{I \to \underline{I}} E(\eta_{id} | \eta_{id} \geq -\underline{I}) = -\underline{I} \tag{12}$$

From, equation 7, this implies that: $(\overline{NU} + x_{id}\delta)E(\widehat{Z_{id}|T_{id}} = 1) - c_{id} = \tau_d$ among patients with $I_{id} = \underline{I}$. Thus, we proceed as follows. We identify marginal patients as patients in the lowest 5 percentiles of I_{id}. We compute $(\overline{NU} + x_{id}\delta)E(\widehat{Z_{id}|T_{id}} = 1) - c_{id}$ for those patients, which gives us estimates $\hat{\tau}_d^* = \tau_d + v_{id}$ of the absolute magnitude of τ for each of the marginal patients. This implies that $\hat{\tau}_d^* = K + \hat{\tau}_d - v_{id}$ so we can then regress $\hat{\tau}_d^*$ on the estimated $\hat{\tau}_d$ to recover the constant K which allows us to appropriately normalize τ and determine for which doctors $\tau_d < 0$.

6.3 Structural Estimation

We use a generalized method of moments estimator to estimate the structural model. The testing equation 5 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:

$$L(\beta, g) = \sum_i [t_i \ln g(X_i'\beta) + (1 - t_i)(1 - \ln g(X_i'\beta))] \tag{13}$$

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:

$$\hat{g}_{-i,d} = \frac{\sum_{j \neq i} k\left(\frac{(X_j - X_i)'\beta}{h}\right) t_i}{\sum_{j \neq i} k\left(\frac{(X_j - X_i)'\beta}{h}\right)} \tag{14}$$

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 $X_i'\beta$.

Because of the large number of fixed effects in the model (over 7,000), it is infeasible to simultaneously estimate all parameters. Instead, we split the sample into 10 subsamples and estimate each

subsample individually. To make sure our estimates are comparable across samples, we include the five doctors with the most patients in every subsample. The doctor with the most patients provides the normalization $\theta_1 = 0$ and in all samples after the first the relative values of doctors 2-5 are fixed so that the normalization remains the same.

We construct moments from the first order condition of the likelihood function in equation 13 with $\hat{g}$ substituted for g. Additional moments are constructed from the regression equation 7. In particular, for each regressor, we construct $\frac{1}{N}\sum_i x_i(y_i - x_i\beta)$ where the regressors include the doctor fixed effects normalized by $\hat{NU}_{id} + \tilde{x}_{id}\delta$. Finally, we impose the additional constraint that $\theta_d = \hat{NU}_{id}\alpha_d - \tau_d$.

6.4 Structural Results

Figure 2 shows the relationship between the underlying estimated propensity to be tested and the observed probability of testing. As we might expect, this function is convex for large values: for most patients a single warning sign is not worrying, but in the presence of several other warning signs the marginal impact on the likelihood of testing increases.

Table 4 reports the marginal effects from estimation of the testing equation. Column (2) shows the coefficients from a linear probability model in which testing is regressed on covariates along with the standard errors of those estimates. The two sets of estimates are very similar. Older patients are substantially less likely to be tested. Black and hispanic patients are less likely to be tested. Patients who have had a pulmonary embolism in the past are 2 percentage points more likely to be tested (compared to a mean of 3.6% tested in our data). Likewise, several other comorbidity indices we include predict increased testing.

Next, we consider the distribution of τ_d resulting from estimation of equation 7. The distribution of the resulting raw τ_d is shown in Figure 3. These initial estimates imply that 80% of doctors in our sample are in this sense "overtesters". The distribution is non-normal because many of the overtesters have 1 or 0 observed positive tests. This means their estimated τ is substantially less than 0, but measured imprecisely. The apparently missing mass between 0 and -3000 reflects the fact that a small in magnitude by negative τ is only possible for doctors with a very large number of tests since it can only result from a non-zero number of positive tests which is nonetheless a small fraction of overall tests. This point underscores the need to correct for the variance in τ when constructing the empirical distribution.

We do this using the "empirical Bayes" techniques. That is, we assume that $\hat{\tau}_d = \tau_d + v_d$ where $\hat{\tau}_d$ gives our estimated τ and τ_d the true value. In this framework, the best linear predictor of the fixed effect τ_d (which is also an estimate of the posterior mean under normality) is given by:

$$\tau_{EB} = \frac{Var(\tau_d^{EB})\hat{\tau}_d}{Var(\tau_d^{EB}) + Var(v_d)} + \frac{Var(v_d)\overline{\hat{\tau}}_d}{Var(\tau_d^{EB}) + Var(v_d)} \tag{15}$$

where $\overline{\hat{\tau}}_d$ is the mean of the estimated values. We estimate the appropriate scaling factor via random effects estimation of equation 7. The resulting distribution of τ^{EB} is graphed in Figure 4. The adjusted τs are all less than 0 and nearly all lie between -$400 and -$900. Taking the adjusted τs literally would imply that everyone in our sample is an overtester. This conclusion is too strong: because positive tests are so rare, we have only imprecise estimates for each doctor; with more information, the posterior distribution in figure 4 would be more diffuse. Nonetheless, this analysis suggests that at least 80% of doctors are overtesters.

Table 5 replicates our reduced form analysis in a structural setting. We regress the estimated τ_{EB} parameters on the potential determinants of inefficient testing. We again find that more experienced doctors and doctors in lower spending regions are less likely to overtest. A 10-year increase in doctor experience is associated with an $80 increase in the testing threshold (where a larger value is associated with less overtesting). A 10% increase in regional spending correlates with a $170 decline in the testing thresholds - so doctors in higher spending regions are more likely to overtest. Column (1) of table 5 uses the full sample of doctors and column (2) considers only emergency room doctors - there is little difference between the two samples (Column (3) is discussed in the next section).

7 Robustness of Structural Model

7.1 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 in fact be tests performed for a different indication.

There are four main alternative indications for CT scans in an emergency room setting: trauma,

lung or chest cancers, pleural effusion, and pneumonia. In the case of trauma, pleural effusion, and cancer workups, we distinguish these indications on the basis of diagnosis codes recorded on the same day as the ER evaluation. In case of pneumonia, we demonstrate the robustness of our structural model to considering this alternative diagnosis as an indication of a "positive" test result.

Trauma patients are excluded by omitting any patients with a diagnosis code associated with injury or fracture. 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. In our Medicare sample, the fraction of total chest CTs performed on trauma patients is 17%, and we exclude these patients from our analysis.

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

CT scans are more commonly used to aid in the treatment of pleural effusion by guiding a procedure rather than to make the initial diagnosis. As a result, we exclude patients with pleural effusion from the sample, since their chest CTs cannot be valued using the diagnostic testing framework outlined here.

Pneumonia diagnoses present a more complex threat to the analysis, since these patients cannot be consistently excluded from the sample due to the nonspecific symptoms and diversity of patient history that may suggest workup for these diseases. Pneumonia is more commonly diagnosed with a chest x-ray, but physicians may choose to supplement or substitute the x-ray with the more detailed data from a CT scan.

For this reason, we consider an extension of the above model in which multiple outcomes are permitted. We consider this extension in the simplest case of the model, where $\delta = 0$ so no heterogeneity in NU is permitted. More precisely, suppose there are k possible outcomes which can be detected by the CT scan. Then we can write the doctors decision of whether or not to test as given by:

$$\sum_k q_{id}^k NU_k - c_{id} \geq \tau_d \tag{16}$$

where q_{id}^k is the probability of a positive test for condition k and is given by:

$$q_{id}^k = x_{id}^k \beta + \alpha_d^k + \eta_{id}^k \tag{17}$$

We show in Appendix B that this implies we can recover τ from a regression of the indicators for a positive test for each condition weighted by the utility of a positive test on τ_d, c_{id} and an appropriately defined testing propensity.

$$\sum_k NU_k E(Z_{id}^k | T_{id} = 1) = \tau_d + c_{id} + I_{id} + g(I_{id}) \tag{18}$$

In particular, we estimate this equation allowing for pneumonia as an alternative positive test. As with pulmonary embolism, positive tests are identified using inpatient diagnosis codes amongst patients who are admitted to the hospital following their emergency room evaluation. We assume that the net utility associated with using a chest CT to diagnose pneumonia is bounded by the cost of a chest x-ray, which could alternatively have been used to make the diagnosis. Chest x-rays are reimbursed at around $30 per scan in the Medicare claims data. The results in Column 3 of Table 5 show that allowing for heterogeneity in pneumonia diagnoses across doctors has little impact on our conclusions.

7.2 Differences in Doctor Discernment

The model so far has assumed that doctors differ only in the cost-benefit threshold τ_d they use to determine whether to test. Specifically, we have assumed that η_{id} is i.i.d. across doctors which implies that doctors do not systematically differ in their ability to recognize patients who will test positive given unobservable factors.

We can relax this assumption by considering a parametric version of our model in which η_{id} is allowed to vary across doctors. Intuitively, some doctors may have no additional information beyond what is observable to the econometrician. For these doctors, $Var(\eta_{id}) = 0$. At the other extreme, some doctors may know with near certainty which patients will test positive - for these doctors, q_{id} $0, 1$ and η_{id} has larger variance.

We want the model to allow for the possibility that doctors with more discernment (understood as a higher variance in η) test less. For this to be the case with a normally distributed η, it must

be that many patients would be tested were $\eta = 0$. To allow for this possibility, we add a bernoulli term $v_{id} = A$ with probability p to the model. We can think of this term as something like "chest pain reported" which is observable to *all* doctors makes it more likely that patients will be tested, but is not observable to the econometrician. The model is thus exactly like the one in the previous section with η_{id} replaced by $\eta_{id} + v_{id}$ and relaxing the assumption that η_{id} is homoskedastic.

The estimating equations for this model are given in Appendix A. We are currently working on estimating this model variation; results are not reported in the current draft.

7.3 Misweighting Observable Characteristics

The model above also assumes that the only "error" doctors can make is to test patients if the costs exceed the benefits. (Of course, this 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).

In this section, we consider the alternative possibility that doctor's systematically misweight observable characteristics in deciding whether to test. That is, assume that doctors' belief about the probability of a positive test is given by:

$$q'_{id} = x_{id}\beta' + \alpha_d + \eta_{id} \tag{19}$$

while the actual probability remains:

$$q_{id} = x_{id}\beta + \alpha_d + \eta_{id} \tag{20}$$

With this change, the derivation of the model in Section 5 continues to hold, except for the assumption that doctors have rational expectations. That is, it is no longer the case that $E(q'_{id}|T_{id} = 1) = E(q_{id}|T_{id} = 1) = E(Z_{id}|T_{id} = 1)$. Instead we have: $E(Z_{id}|T_{id} = 1) = E(q'_{id}|T_{id} = 1) + x_{id}(\beta - \beta')$ which yields the equation:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c_{id}}{\overline{NU} + \tilde{x}_{id}\delta} + x_{id}(\beta - \beta') + I_{id} + g(I_{id}) \tag{21}$$

This equation allows us to recover τ_d *and* $\beta - \beta'$. So we can test both whether doctors are overtesting in the threshold sense or testing the wrong patients because they misweight patient characteristics. Intuitively, non-zero coefficients on the x's imply that they 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.

Estimation of this variation on the model is forthcoming, but results are not reported in this draft.

8 Simulations and Welfare

8.1 Welfare Cost of Overtesting

Given the estimated taus, we can simulate how testing behavior would differ if all doctors tested only when expected benefits exceeded costs. To perform this simulation, we must first determine the relative magnitude of τ_{EB} and the other variables included in our testing model. This relationship is not identified from what we have estimated so far: τ_{EB} is expressed in dollars, while the variables in the testing equation are in units of whatever normalization was imposed in that equation (which in our case was the impact on testing of being in age bracket 80-85). To determine the appropriate scaling of τ_{EB}, we re-estimate the structural model directly including our empirical Bayes estimate τ_{EB} as a variable in our testing equation with coefficient normalized to -1.[4] This allows us to re-estimate all of the other parameter values in units of τ_{EB}.]

We simulate how testing behavior would change if no doctors overtested. We find that the fraction of patients given a chest CT in our sample would drop from 3.63% to 3.08%. As shown in Table 6, the total dollar spending on CT scans - including both the financial cost of the test and the cost of admitting patients who tested positive - would fall by 11.4%. The medical benefits would increase by 27% due to the fact that a larger ratio of true positives to false positives would substantially increase the value of treatment conditional on testing relative to the medical hazards posed by treatment. Together, these factors imply that the total net benefits of testing would nearly double.

[4]The empirical Bayes values are required here because we are effectively putting τ_d on the right-hand side of an estimating equation, so the coefficient on the unadjusted values would be severely biased due to measurement error. In a linear model, the empirical Bayes measurement error correction would be exact.

9 Conclusion

While it is commonly believed that the 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. Constructing public policy to reduce wasteful spending requires us to first identify instances of overspending, and second, to identify the conditions driving the overuse behavior. This paper works to bridge this gap by precisely estimating the amount of wasteful spending in one specific context, emergency room CTs to diagnose pulmonary embolism, and then exploring the determinants of that variation in wasteful spending across physicians, regions, and hospitals.

By estimating a structural model of physician testing behavior, we find that 80% of doctors evaluating emergency room 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. Less experienced physicians and those practicing in high-spending regions (as measured by the Dartmouth Atlas) are more likely to perform wasteful tests. If all doctors adopted the optimal testing strategy, testing only when expected benefits exceed expected costs, 15% fewer chest CT scans would be performed and the welfare associated with CT tests for pulmonary embolism would roughly double.

These findings provide support for the hypothesis that overuse of medical services despite negative net benefits is a pervasive driver of health care spending. By measuring physician-level preferences for under- or over-testing, we are able to further 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 his propensity to over-test. More generally, the doctor-specific measure of overtesting we develop can serve as a "left-hand side" variable in any analysis seeking to understand the determinants of efficient medical care.

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.

Doyle, J. (2007). Returns to local-area health care spending: using health shocks to patients far from home. Technical report, National Bureau of Economic Research.

Doyle, J., J. Graves, J. Gruber, and S. Kleiner (2012). Do high-cost hospitals deliver better care? evidence from ambulance referral patterns. Technical report, National Bureau of Economic Research.

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

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.

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.

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.

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	76.7
Female	0.59	0.59	0.59
White	0.86	0.89	0.90
Black	0.10	0.08	0.09
History of PE	0.003	0.006	0.02
Doctor, hospital and region characteristics			
Physician experience	16.6	16.0	16.6
	(9.0)	(9.0)	(9.0)
HRR avg spending (in $)	57,040	56,880	56,041
	(7380)	(7220)	(7070)
Academic hospital	0.36	0.35	0.36
Top 50 research med. school	0.31	0.3	0.31
Top 50 primary med. school	0.29	0.28	0.29
No. of observations	6,119,406	133,878	8,604

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: Reduced form relationship between positive tests and doctor characteristics

	Dependent variable: positive chest CT scan					
Independent variables	(1)	(2)	(3)	(4)	(5)	(6)
Doctor experience (in years)	0.00050**	0.00054**	0.00054**	0.00054**	0.00052**	0.00052**
	(0.00008)	(0.00008)	(0.00008)	(0.00008)	(0.00009)	(0.00009)
Top 50 research med school	0.00277	0.00384	0.00311	0.00354	0.00356	0.00387
	(0.00220)	(0.00220)	(0.00210)	(0.00207)	(0.00241)	(0.00240)
Top 50 primary care med. school	-0.00004	-0.00136	0.00045	-0.00056	0.00341	-0.0003
	(0.00222)	(0.00221)	(0.00212)	(0.00210)	(0.00240)	(0.00238)
Female doctor					-0.00245	-0.00251
					(0.00149)	(0.00148)
Academic hospital	0.00152	-0.00020	-.000082	-0.00222	-.00046	-0.00173
	(0.00179)	(0.00177)	(0.00168)	(0.00168)	(0.00189)	(0.00191)
For profit hospital	-0.00453	-0.00450*	-0.00422	-0.00496*	-0.00580*	-0.00696**
	(0.00244)	(0.00248)	(0.00240)	(0.00235)	(0.00257)	(0.00258)
Log(avg HRR spend. per benef.)	-0.05544**	-0.04239**	-0.05819**	-0.05399**	-0.05229**	-0.05230**
	(0.00940)	(0.01242)	(0.00893)	(0.01218)	(0.00923)	(0.01255)
Per capita income (in thousands)	0.00008	0.00033*	0.00000	0.00035*	0.00000	0.00000
	(0.00016)	(0.00017)	(0.00016)	(0.00016)	(0.00016)	(0.00017)
Malpractice damage caps	-0.00218	0.00261	-0.00219	0.00323	-0.00246	0.00325
	(0.00141)	(0.00424)	(0.00211)	(0.00340)	(0.00236)	(0.00417)
Joint & several liability reform	0.00164	0.01084*	0.00200	0.00969*	0.00215	0.0088
	(0.00211)	(0.00424)	(0.00217)	(0.00458)	(0.00239)	(0.00499)
Controls for comorbidities	No	No	Yes	Yes	Yes	Yes
Region fixed effects	None	State	None	State	None	State
Physician sample	All doctors	All doctors	All doctors	All doctors	EM doctors	EM doctors
No. of observations	144,244	142,487	142,487	142,487	116,603	115,176
No. of doctors	32,921	32,921	32,921	32,921	24,273	24,273

Notes: Table reports results from 6 reduced form regressions of whether a patient receives a positive test on physician, region, and hospital characteristics, and patient control variables. An observation is a patient tested with a chest CT scan within one day of a submitted emergency room bill. All regressions include controls for patients race, sex, and one-year age bins. Standard errors are clustered at the hospital referral region level. Data is from the Medicare claims 2000-2009, the American Hospital Association annual survey, the American Medical Association masterfile, and the Dartmouth Atlas. ** denotes statistical significance at the 1% level; * at the 5% level.
Even numbered columns include state fixed effects.
Columns 3 through 6 also include controls for Elixhauser comorbidities and pulmonary embolism risk factors.
Columns 5 and 6 restrict to patients who are evaluated by a physician who specailizes in emergency medicine.

Table 3: Calibrating the model

Parameter	*Value*	*Definition*	*Source*
s	0.83	test sensitivity	Lesler et al., 2009
fp	0.05	false positive	Stein et al., 2006
MB_{id}	0.025*VSL*	medical benefit of testing	Lesler et al., 2009
MC_{id}	0.0017*VSL*	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
VSL	1,500,000*	value of a statistical life	Murphy & Topel, 2006

*We allow VSL to vary with age according to the schedule in Murphy & Topel (2006). It is $1.5 million for a 75 year-old, and declines by approximately $100,000 per year.

Table 4: Estimates of the testing equation for the structural model

	Dependent variable: Chest CT test	
Independent variables: Patient characteristics	(1)	(2)
Age 70-74	-0.0003	-0.0008
		(0.0005)
Age 75-79	-0.0022	-0.0029**
		(0.0005)
Age 80-84	-0.0050	-0.0047**
		(0.0005)
Age 85-89	-0.0045	-0.0068**
		(0.0005)
Age 90-94	-0.0095	-0.0119**
		(0.0006)
Age 95-99	-0.0195	-0.0165**
		(0.0009)
Black	-0.0151	-0.0130**
		(0.0038)
Hispanic	-0.0111	-0.0082**
		(0.0005)
Asian	-0.0030	-0.0003
		(0.0016)
Native American	-0.0023	0.0003
		(0.0015)
Other race	-0.0069	-0.0061**
		(0.0012)
Unkown race	-0.0044	-0.0023
		(0.0028)
Female	0.0029	0.0032**
		(0.0003)
History of pulmonary embolism	0.0205	0.0283**
		(0.0023)
History of thoracic aortic disection	0.0106	0.0114**
		(0.0019)
History of other aortic disection	0.0081	0.0122**
		(0.0027)
History of deep vein thrombosis	0.0055	0.0035**
		(0.0008)
Previously admitted within 30 days	0.0044	0.0036**
		(0.0008)
Previously admitted within 7 days	0.0107	0.0113**
		(0.0012)

Notes: Table reports results from structural model (column 1) and an OLS regression (column 2) of whether an ER patient is evaluated with a chest CT on a vector of patient characteristics. Patients are excluded if their evaluating physician ordered fewer than 10 CT scans in the full sample, or fewer than 4 CT scans after imposing exclusions (see section 2.2). Observation is a patient evaluated in the ER; there are 2,010,951 ER evaluations from 6828 doctors. Standard errors are in parentheses. ** denotes statistical significance at the 1% level; * at the 5% level.

Table 5: Regressions of testing threshold on physician characteristics and practice environment

	Dependent variable: physician's testing threshold		
Independent variables	(1)	(2)	(3)
Doctor experience (in years)	8.79**	7.28	7.99**
	(3.53)	(3.85)	(3.53)
Log(avg HRR spend. per benef.)	-1746**	-1447**	-1731**
	(243)	(263)	(243)
Academic hospital	60	61	58
	(58)	(63)	(58)
Top 50 research med school	96	29	95
	(85)	(89)	(85)
Top 50 primary care med. school	58	-27	-58
	(87)	(92)	(87)
Adjust for pneumonia diagnoses?	No	No	Yes
Physician sample	All doctors	EM doctors	All doctors

Notes: Table reports results from 3 separate regressions of a physician's testing threshold on his experience, regional spending, practice environment, and training. The testing thresholds are estimated from the structural model outlined in Section 4. There are 6828 physician observations. Standard errors are in parentheses. ** denotes statistical significance at the 1% level; * at the 5% level.

Column 1 reports results from the baseline specification described in the text over the full sample of patients and doctors.

Column 2 restricts the sample to emergency medicine specialized physicians.

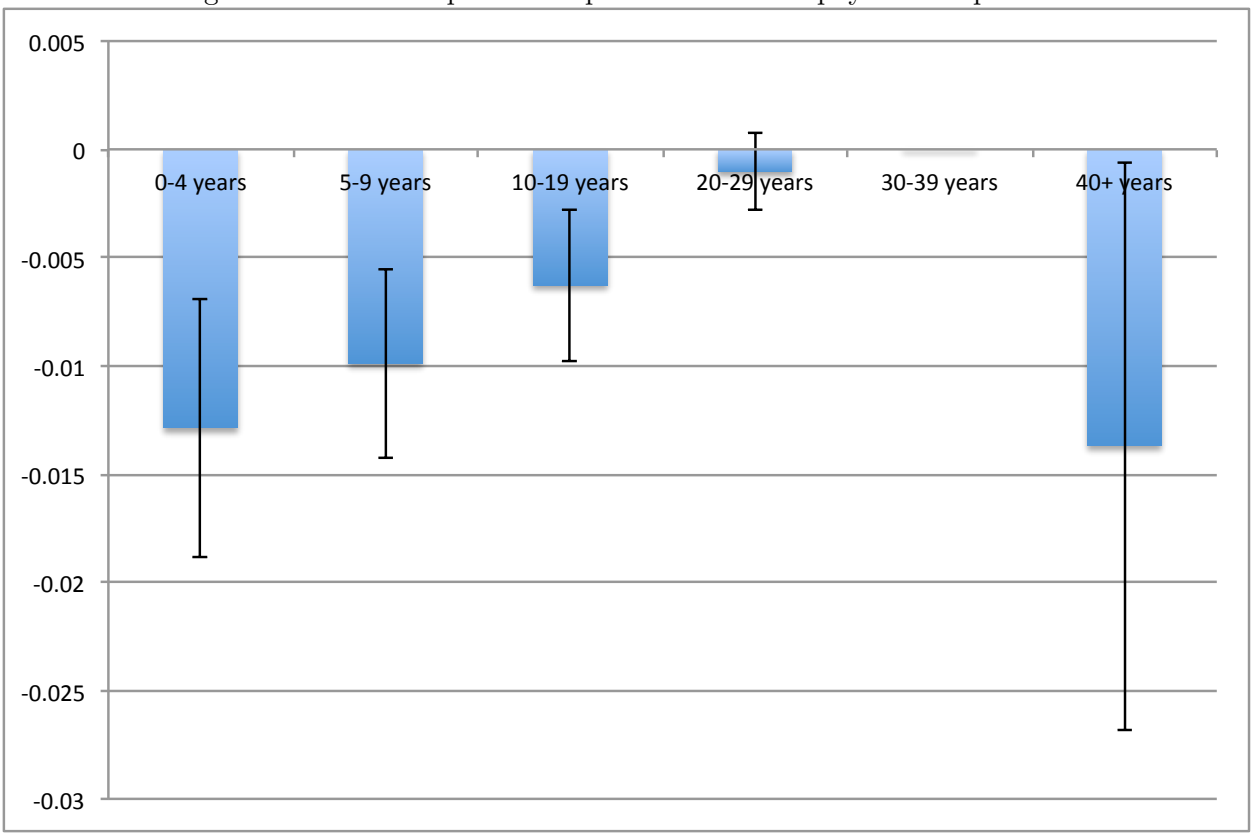
Column 3 accountes for the value of chest CT scans in diagnosing pneumonia.

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

Welfare metric	*Actual testing behavior* (1)	*Simulated behavior with no over-testing* (2)
Percent of patients tested with chest CT	0.0364	0.0308
Number of patients tested with chest CT	73079	62039
Total financial costs of testing (millions)	36.1	32.0
Total medical benefits of testing (millions)	55.2	70.2
Net benefits of testing (millions)	19.1	38.2
Costs per test	494.5	437.9
Benefits per test	756	1131.1
Net benefits per test	261.6	615.3

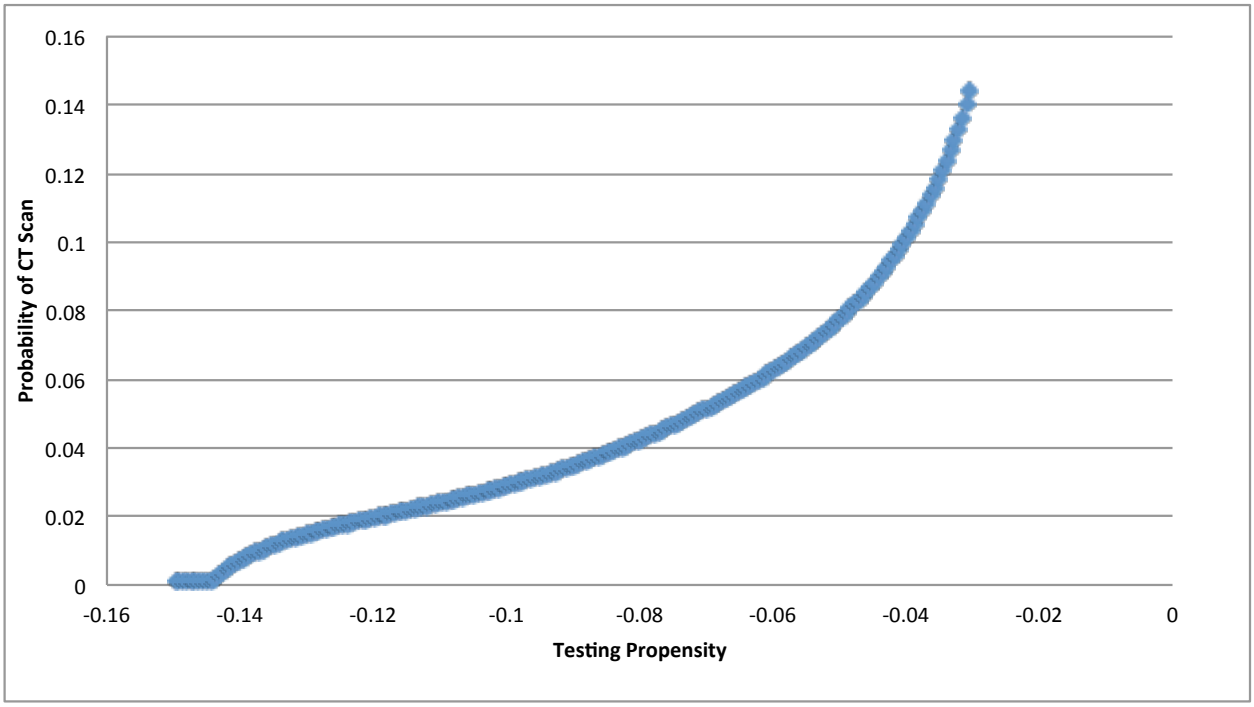
Notes: This table presents results from the welfare simulations detailed in Section 8. Column 1 describes testing behavior and outcomes given physician's observed testing thresholds. Column 2 presents simulated results estimating testing behavior and benefits in a counterfactual world in which no physician over-tests.

Figure 1: Relationship between positive tests and physician experience



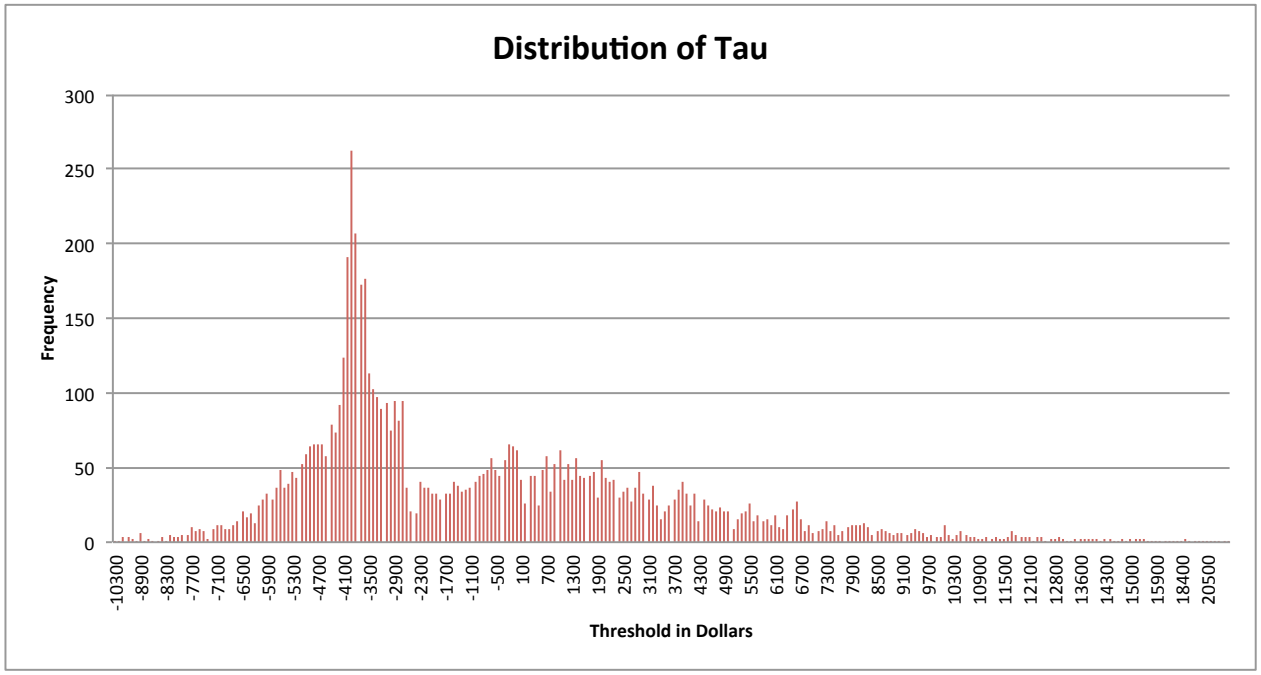
Notes: This figure plots coefficients from a regression of whether the patient tested positive for PE on physician experience bins, controlling for patient age, race, sex, comorbidities, risk factors for pulmonary embolism, state fixed effects, as well as regional Medicare spending, teaching hospital status, and physician medical school quality. The 30-39 year experience group is the omitted category and thus normalized to zero. Error bars represent the 95% confidence interval. An observation is a patient who receives a chest CT within one day of an emergency room evaluation. There are 142,487 observations.

Figure 2: Relationship between estimated testing propensity and probability of testing



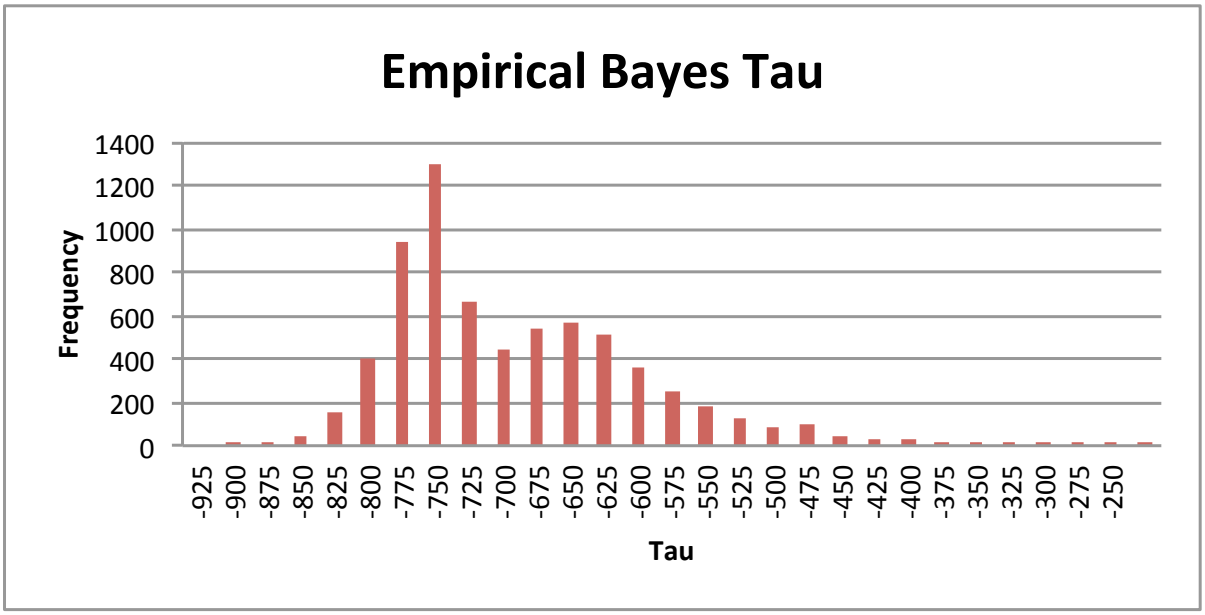
Notes: This figure plots the predicted testing propensity estimated by equation (4) and reported in Table 4, column 1, on the x-axis against the probability that the patient receives a CT scan.

Figure 3: Histogram of estimated testing thresholds, τ



Notes: This figure plots a histogram of the values of tau, the physician's testing threshold, estimated from the structural model, following equation 6. Results are plotted for each of the 6,828 physicians in the sample.

Figure 4: Histogram of Bayesian adjusted estimated testing thresholds, τ



Notes: This figure plots a histogram of the values of **τ,** the physician's testing threshold, *after applying the Bayesian shrinkage estimator.* Results are plotted for each of the 6,828 physicians in the sample.

A Heteroscedastic Model (differential discernment)

We want the model to allow for the possibility that doctors with more discernment (understood as a higher variance in η) test less. For this to be the case with a normally distributed η, it must be that many patients would be tested were $\eta = 0$. To allow for this possibility, we add a bernoulli term $v_{id} = A$ w/ probability p to the model. We can think of this term as something like "chest pain reported" which is observable to all doctors makes it more likely that patients will be tested but is not observable to the econometrician. The model is thus exactly like the one in the previous section with η_{id} replaced by $\eta_{id} + v_{id}$ and relaxing the assumption that η_{id} is homoskedastic.

Now, rewriting, we know that a doctor will test a patient if:

$$x_{id}\beta + \frac{\theta_d + \alpha_d x_{id}\delta - \hat{c}_{id}}{\widehat{NU_{id}} + x_{id}\delta} + v_{id} + \eta_{id} \geq 0 \tag{22}$$

So the testing probability becomes:

$$\begin{aligned} P(T_{id} = 1) &= P(v_{id} + \eta_{id} \geq -I_{id}) \\ &= pP(\eta_{id} \geq -I_{id} - A) + (1-p)P(\eta_{id} \geq -I_{id}) \\ &= p\Phi\left(\frac{-I_{id} - A}{\sigma_\eta^2(d)}\right) + (1-p)\Phi\left(\frac{-I_{id}}{\sigma_\eta^2(d)}\right) \end{aligned} \tag{23}$$

Following exactly the same steps as the previous section gives:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + \hat{c}_{id}}{\widehat{NU_{id}} + x_{id}\delta} + I_{id} + g(I_{id}, d) \tag{24}$$

where $g(I_{id}, d) = E(\eta_{id} + v_{id}| - (\eta_{id} + v_{id}) \leq I_{id})$. We can write this as:

$$\begin{aligned} g(I_{id}, d) &= E(\eta_{id} + v_{id}| - (\eta_{id} + v_{id}) \leq I_{id}) \\ &= pE(\eta_{id} + A| - (\eta_{id} + A) \leq I_{id}) + (1-p)E(\eta_{id}| - \eta_{id} \leq I_{id}) \\ &= pA + pE(\eta_{id}|\eta_{id} \geq -I_{id} - A) + (1-p)E(\eta_{id}|\eta_{id} \geq -I_{id}) \\ &= pA + p\frac{\phi\left(\frac{-I_{id}-A}{\sigma(d)}\right)}{1 - \Phi\left(\frac{-I_{id}-A}{\sigma(d)}\right)}\sigma(d) + (1-p)\frac{\phi\left(\frac{-I_{id}}{\sigma(d)}\right)}{1 - \Phi\left(\frac{-I_{id}}{\sigma(d)}\right)}\sigma(d) \end{aligned} \tag{25}$$

B Testing with Multiple Outcomes

Suppose there are k possible outcomes which can be detected by the CT scan. Then we can write the doctor's decision of whether or not to test as given by:

$$\sum_k q_{id}^k NU_k - c_{id} \geq \tau_d \tag{26}$$

where q_{id}^k is the probability of a positive test for condition k and is given by:

$$q_{id}^k = x_{id}^k \beta + \alpha_d^k + \eta_{id}^k \tag{27}$$

Define $\theta_d = \sum_k NU_k \alpha_d^k - \tau_d$. Plugging our specifications for the probability of a positive test into the testing equation yields:

$$\sum_k NU_k x_{id}^k \beta + \theta_d - c_{id} + \sum_k NU_k \eta_{id}^k \geq 0 \tag{28}$$

As above define: $I_{id} \equiv \sum_k NU_k x_{id}^k \beta + \theta_d - c_{id}$.[5]. From equation 28, we can compute the expected benefits conditional on testing, which are given by:

$$E(B_{id}|T_{id} = 1) = \tau_d + I_{id} + g(I_{id}) \tag{29}$$

where $g(I_{id}) = E(\tilde{\eta}_{id}| - \tilde{\eta}_{id} \leq I_{id})$ is an (unknown) function of I_{id} and $\tilde{\eta}_{id} = \sum_k NU_k \eta_{id}^k$.

Let Z_{id}^k be an indicator for whether a test for condition k is positive or negative. If doctors have rational expectations, we must have $E(q_{id}^k|T_{id} = 1) = E(Z_{id}^k|T_{id} = 1)$. Given these rational expectations and equation 3, we can write the expected benefits as $E(B_{id}|T_{id} = 1) = \sum_k NU_k E(Z_{id}^k|T_{id} = 1) - c_{id}$. Plugging this into equation 6 and rearranging yields:

$$\sum_k NU_k E(Z_{id}^k|T_{id} = 1) = \tau_d + c_{id} + I_{id} + g(I_{id}) \tag{30}$$

[5]Note that in the single outcome case, we normalized the testing equation by NU to eliminate heteroskedasticity. In this case, it is more convenient to keep NU in the testing equation and in the propensity I_{id} - this normalization is the reason the equations outlined here with $k = 1$ do not match exactly with the equations in the single-outcome case with $\delta = 0$