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

Jason Abaluck and Leila Agha
with Chris Kabral, Ali Raja, Jay Schuur, Arjun Venkatesh
and maybe a few other doctors?*

September 5, 2012

Abstract

Studies have documented 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 Nathan Hendren, Mitch Hoffman, Danielle Li, David Molitor and a couple of people at hospitals in Boston who helped with the data whose names we will look up in our emails and include in the next draft. Funding for this work was provided by ...
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 \textit{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.

We build on the theoretical insights made by Chandra and Staiger (2011) (hereafter CS) who use a structural model to estimate the medical returns to heart attack treatment and document variation in utilization due to 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 $\tau_d$ and in structural estimation they seek to recover $\tau_d$. 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 \textit{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%.

We end by examining the external validity of our estimates using the relationship between regional spending and regional estimates of wasteful and non-wasteful CT spending. We consider a model in which the magnitude of the coefficient on wasteful CT spending controlling for non-wasteful CT spending (“need”) tells us whether other services in the health economy have a high fraction of waste. The underlying assumption of the model is that wasteful spending is correlated across services within regions so that “waste predicts waste”—provided this assumption holds, the model can be used to determine how the proportion of spending which is wasteful in the health economy as a whole relates to the proportion of spending on CT scans. In the simplest possible case of our model, a coefficient much larger than 1 on wasteful CT spending implies that for each dollar of wasteful CT spending, there are several dollars of other wasteful spending in the economy. Using this method, we estimate that more than 50% of all health spending is wasteful.

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 the threshold $\tau_d$ described above which
indicates whether or not a doctor is an overtester. Section 5 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, ranging from around 2.5% within three months for a mild PE (Lessler, Isserman, Agarwal, Palevsky, and Pines 2010) to as high as 30% for severe PE (Thompson and Hales 2012).

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 comorbidities. 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,

---

1Conditions 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.
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 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.

### 3.3 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 Model of Testing Behavior

We outline 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 $\tau_d$ denote this threshold. Then we assume that doctors test if and only if $B_{id} \geq \tau_d$. If $\tau_d$ 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 $\tau_d$ 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 $\tau_d$ 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. While Chandra and Staiger’s model focuses on identifying overuse of medical treatments, we have adapted the framework to analyze overuse of medical tests. A key advantage of investigating the efficient use of medical testing is that doctor threshold parameters, $\tau_d$, 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—that of 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.

Given this assumption that negative tests are not ex post medically valuable, the net benefits of testing are given by 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 imply that doctor \( d \) tests patient \( i \) if and only if:

\[
q_{id}NU_{id} - c_{id} \geq \tau_d
\]  

(1)

We assume that the probability of a positive test is given by:

\[
q_{id} = x_{id}\beta + \alpha_d + \eta_{id}
\]  

(2)

where \( x_{id} \) are observed patient characteristics, \( \alpha_d \) are doctor level fixed effects, and \( \eta_{id} \) are factors observable to the doctor but unobservable to the econometrician which impact the likelihood that a test is positive. For example, \( \eta_{id} \) might include symptoms reported by the patient such as chest pain. For now, we assume that \( \eta_{id} \) is i.i.d. across doctors and patients.\(^2\)

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

\[
NU_{id} = \overline{NU}_{id} + \tilde{x}_{id}\delta
\]  

(3)

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.

\[\text{Define } \theta_d = \overline{NU}_{id}\alpha_d - \tau_d.\]  

Plugging our specifications for the probability of a positive test and

\(^2\)This is all we do in this draft. We can in principle estimate a model with heteroskedastic \( \eta \) provided we make a parametric assumption about \( \eta_{id} \) (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.
for net utility into the testing equation and rearranging yields the final form of the testing equation:

$$x_{id}\beta + \frac{\theta_d + \alpha_d\bar{x}_id\delta - \bar{c}_id}{\bar{NU}_{id} + \bar{x}_id\delta} + \eta_{id} \geq 0 \ (4)$$

This yields a standard semiparametric binary choice model of testing. We next show how the threshold parameters $\tau_d$ 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 propensity by $I_{id} \equiv x_{id}\beta + \frac{\theta_d + \alpha_d\bar{x}_id\delta - \bar{c}_id}{\bar{NU}_{id} + \bar{x}_id\delta}$. From equation 4, we can compute the expected benefits conditional on testing, which are given by:

$$E(B_{id}|T_{id} = 1) = \tau_d + (\bar{NU} + \bar{x}_id\delta)I_{id} + (\bar{NU} + \bar{x}_id\delta)g(I_{id}) \ (5)$$

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 1, we can write the expected benefits as $E(B_{id}|T_{id} = 1) = (\bar{NU} + \bar{x}_id\delta)E(Z_{id}|T_{id} = 1) - c_{id}$. Plugging this into equation 5 and rearranging yields:

$$E(Z_{id}|T_{id} = 1) = \frac{\tau_d + c_{id}}{\bar{NU} + \bar{x}_id\delta} + I_{id} + g(I_{id}) \ (6)$$

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

Intuitively, the propensity to test is determined by two factors: the expected benefits of testing and doctor’s testing threshold. Thus, if expected benefits vary conditional on the propensity to test, this must be because doctors have different thresholds for the expected net benefits of testing at which they are just indifferent between testing and not testing.

Further, 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. So 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 18 only allows the identification of testing thresholds $\tau_d$ up to a constant normalization. From this equation, we cannot directly recover the absolute magnitudes of $\tau_d$—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 $\tau_d$ and thus determine which physicians are over-testing and which are under-testing.

Equation 18 motivates the natural reduced form exercise of regressing the indicator 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 $\tau_d$ thresholds on covariates scaled by $NU$. In the next section, we provide estimates from this reduced form regression and in Section 6, we estimate the full structural model to recover the $\tau_d$ parameters which we then use to understand the determinants of overtesting as well as the welfare costs and how this relates to the overall level of medical expenditures.

5 Reduced Form Estimation

5.1 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

\[^3\text{A somewhat subtle point is that only because of heterogeneity in }\tau_d\text{ can we separately identify heterogeneity in net utility (which scales with }\alpha_d\text{) from non-linearities in the function relaxing the propensity to consume to the probability of testing. Thus, }\delta\text{ is only separately identified when equations 5 and 18 are estimated jointly.}\]
regressions are estimated over the sample of tested patients, and they take the following form:

\[ Z_{id} = a_1 + a_2 Y_d + a_3 X_i + \epsilon_id \]  

(7)

\( Z_{id} \) 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 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 allowing doctors to face different distributions of unobservable patient characteristics [NOT IN THIS DRAFT!].

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.

5.2 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 all
physicians and years, along with the richest set of controls including both 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.

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.

6 Calibration and Estimation of the Structural Model

6.1 Calibration of Parameters

To estimate the model laid out in Section 4, 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).

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:

$$NU_{id} = P(PE_{id}|Z_{id} = 1)(MB_{id} - MC_{id}) + (1 - P(PE_{id}|Z_{id} = 1))(-MC_{id} - CT_{id})$$  (8)

Applying Bayes’ Rule and the law of total probability we can rewrite this in terms of $s$ and $fp$. 
\[
\mathcal{NU} = \frac{s(q_{id} - fp)}{q_{id}(s - fp)}(MB_{id} - MC_{id}) + (1 - \frac{s(q_{id} - fp)}{q_{id}(s - fp)})(-MC_{id} - CT_{id}) \tag{9}
\]

We can therefore write the net benefits of testing as:

\[
B_{id} = q_{id}(\mathcal{NU} + x_{id}\delta) - c_{id}
\]

\[
= \frac{s(q_{id} - fp)}{(s - fp)}(MB_{id} - MC_{id}) + (q_{id} - \frac{s(q_{id} - fp)}{(s - fp)})(-MC_{id} - CT_{id}q_{id} + q_{id}x_{id}\delta - c_{id}) \tag{10}
\]

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

\[
B_{id} = q_{id}(\hat{NU}_{id} + x_{id}\delta) - \hat{c}_{id} \tag{11}
\]

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; 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 the robustness section 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) [NOT IN CURRENT DRAFT].

6.2 Who is an overtester?

As noted in Section 4, equation 5 only allows us to recover the relative values of \(\tau_d\)—we also need to determine an appropriate normalization in order to identify which doctors are overtesters and
which are undertesters. In other words, let $\tau_d^*$ denote the true $\tau$'s and $\hat{\tau}_d$ the $\tau$'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 $\tau_d$ 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 $\tau_d$ which by definition is equal to the expected net benefits for the marginal patient. Formally, note that $\eta_{id}$ is bounded since $q_{id} \in [0, 1]$. Thus, there exists a value $L$ such that, for $I_{id} < L$, patient $i$ cannot be tested. Further, at $L$, we know that $\eta_{id} = \bar{\eta}$. In other words,

$$\lim_{I \to L} g(I_{id}) = \lim_{I \to L} E(\eta_{id} \mid \eta_{id} \geq -L) = -L$$

(12)

From, equation 18, this implies that: $(NU + x_{id}\delta)E(Z_{id} \mid T_{id} = 1) - c_{id} = \tau_d$ among patients with $I_{id} = L$. Thus, we proceed as follows. We identify marginal patients as patients in the lowest 5 percentiles of $I_{id}$. We compute $(NU + x_{id}\delta)E(Z_{id} \mid T_{id} = 1) - c_{id}$ for those patients, which gives us estimates $\hat{\tau}_d^* = \tau_d + v_{id}$ of the absolute magnitude of $\tau$ 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 $\tau$ 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 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:

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

(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 $I_{id} = X_i'\beta$. Specifically, we substitute for $g$ using:
\[
\hat{g}_{-i,d} = \frac{\sum_{j \neq i} k \left( (X_j - X_i)^\prime \beta \right) t_i}{\sum_{j \neq i} k \left( (X_j - X_i)^\prime \beta \right)}
\]

(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 18. 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 \( \tilde{N}\gamma_{id} + \tilde{x}_{id} \delta \). Finally, we impose the additional constraint that \( \theta_d = \tilde{N}\alpha_d - \tau_d \).

### 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 \( \tau_d \) resulting from estimation of equation 18. The distribution of the resulting raw \( \tau_d \) 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 $\tau$ 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 $\tau$ 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 $\tau$ 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 $\tau$ and $\tau_d$ the true value. In this framework, the best linear predictor of the fixed effect $\tau_d$ (which is also an estimate of the posterior mean under normality) is given by:

$$
\tau_{EB} = \frac{V\text{ar}(\tau_d^{EB})\bar{\tau}_d}{V\text{ar}(\tau_d^{EB}) + V\text{ar}(v_d)} + \frac{V\text{ar}(v_d)\bar{\tau}_d}{V\text{ar}(\tau_d^{EB}) + V\text{ar}(v_d)}
$$

(15)

where $\bar{\tau}_d$ is the mean of the estimated values. We estimate the appropriate scaling factor via random effects estimation of equation 18. The resulting distribution of $\tau_{EB}$ is graphed in Figure 4. The adjusted $\tau$s are all less than 0 and nearly all lie between -$400 and -$900. Taking the adjusted $\tau$s 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.

### 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:

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

(16)
where \( q_{id}^k \) is the probability of a positive test for condition \( k \) and is given by:

\[
q_{id}^k = x_{id}^k \beta + \alpha_d^k + \eta_{id}^k
\]  

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

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

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.

8 Simulations and Welfare

8.1 Welfare Cost of Overtesting

Given the estimated \( \tau \)s, we can attempt to determine 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 \( \tau_{EB} \) and the other variables included in our testing model. This relationship is not identified from what we have estimated so far: \( \tau_{EB} \) 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 \( \tau_{EB} \), we re-estimate the structural model directly including our empirical Bayes estimate \( \tau_{EB} \) as a variable in our testing equation with coefficient normalized to
This allows us to re-estimate all of the other parameter values in units of $\tau_{EB}$.

Given our coefficients re-expressed in units of $\tau_{EB}$, 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 PPP, 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.

8.2 Welfare and Spending

Our results so far suggest that 11-12% of spending on CT scans is wasteful, in the sense that if doctors only tested when the expected health benefits from treatment exceed the costs of testing then spending would be lower by that amount. Does this estimate extend to all medical care? The CT spending we observe is only a drop in the bucket of overall medical expenditures - nonetheless, we can use the relationship across regions between overall spending and wasteful and non-wasteful CT spending to examine whether high spending regions are more likely to be high spending because of waste or because they have patients which require more care.

We begin by constructing a measure of wasteful and non-wasteful CT spending in each region. Non-wasteful spending is defined as what doctors would spend with $\tau_{EB} = 0$, and wasteful spending is defined as the difference between the observed CT spending and non-wasteful spending.

The purpose of this section is to develop a simple model of medical spending which allows us to extrapolate the proportion of wasteful spending on all medical care from the proportion of wasteful spending on CT scans combined with the cross sectional relationship between overall spending and wasteful and non-wasteful CT spending. A large coefficient on wasteful CT spending after controlling for non-wasteful CT spending suggests a greater prevalence of waste in other medical services.

We start by assuming that wasteful and non-wasteful spending for service $j$ in region $R$ can be

\[ \text{wasteful spending} = \text{non-wasteful spending} + \text{difference} \]

\[ \text{difference} = \text{observed CT spending} - \text{non-wasteful spending} \]

The empirical Bayes values are required here because we are effectively putting $\tau_d$ 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.
written as:

\[
\begin{align*}
NW_{R,j} &= \xi_R + u_{R,j} \\
W_{R,j} &= \alpha_j(\xi_R + v_R)
\end{align*}
\] (19)

\(\xi_R\) gives expected non-wasteful consumption in each region which we can think of as the “need” for medical care. Any given region may have higher or lower legitimate demand for particular services, which leads to the idiosyncratic error term \(u_{R,j}\).

Wasteful spending in each region depends on the total need for each type of medical service \(\xi_R\), as well as a region-specific degree of waste \(v_R\) that is constant across services. Regions with higher needs for non-wasteful medical care also have more wasteful spending. For example, in the case of CT scans, wasteful spending depends on the number of patients who present with net benefits between \(\tau_d\) and 0. If we scale up the number of patients who present in the emergency room with symptoms related to PE keeping the distribution of net benefits constant, the amount of wasteful spending will increase proportionately. We further assume that the amount of waste for each service scales with \(\alpha_j\), allowing some types of services to be more prone to waste than others in all regions. The model can allow for a few notable special cases: (a) \(\alpha_j = 0\) for all services except chest CT scans \((j = \text{CT})\), so there is no waste in any services besides CT spending; (b) the proportion of waste for all services is the same \((\alpha_j = \alpha_{C\text{T}}\text{ for all } j)\); or (c) cases where other services are more wasteful than CT scans \((\alpha_j > \alpha_{C\text{T}})\).

Summing the above equations over all service types \(j\) within a region \(R\), total spending in each region is given by:

\[
S_R = \sum_j (W_{R,j} + NW_{R,j}) = \xi_R(\sum_j \alpha_j + J) + v_R \sum_j \alpha_j + \sum_j u_{R,j}
\] (20)

We estimate an OLS regression of total spending on the wasteful and non-wasteful components of CT spending in the region. Using the modeling assumptions described above, we can solve for the OLS regression coefficients in this model. That is, we find \(\beta_0\), \(\beta_1\) and \(\beta_2\) such that:

\[
S_R = \beta_0 + \beta_1 W_{R,\text{CT}} + \beta_2 NW_{R,\text{CT}} + \epsilon_{R,j}
\] (21)
with $\text{Cov}(W_{R,CT}, \epsilon_R) = 0$ and $\text{Cov}(NW_{R,CT}, \epsilon_R) = 0$.

Define $\rho = \text{Cov}(NW_{R,CT}, W_{R,CT}) = \alpha_{CT} \text{Var}(\xi)$. Appendix A shows that these assumptions imply:

$$
\begin{align*}
\beta_0 &= E(\xi_R)(\sum_j \alpha_j + J) \\
\beta_1 &= \frac{\sum_j \alpha_j}{\alpha_{CT}} + \frac{(J-1)(\rho \text{Var}(NW_{R,CT}) - \frac{\rho^2}{\alpha_{CT}})}{(\text{Var}(NW_{R,CT}) \text{Var}(W_{R,CT}) - \rho^2)} \\
\beta_2 &= 1 + \frac{(J-1)(\frac{\rho}{\alpha_{CT}} \text{Var}(W_{R,CT}) - \rho^2)}{(\text{Var}(NW_{R,CT}) \text{Var}(W_{R,CT}) - \rho^2)}
\end{align*}
$$

(22)

The variation in non-wasteful spending loads onto both terms, loading more onto whichever term has lower variance (since both wasteful and non-wasteful spending are proportional to medical need). In contrast, the coefficient on wasteful CT spending absorbs the variation in total wasteful spending, since we assume that non-wasteful spending for service $k$ is uncorrelated with wasteful spending for service $j$ after controlling for region-level “need”. In other words, the only reason in our model that regions with high wasteful CT spending would tend to have high spending overall controlling for non-wasteful CT spending is because those regions have higher wasteful spending on other services.

The proportion of wasteful spending in the economy is given by:

$$
\frac{E(\sum_j W_{R,j})}{E(S_R)} = \frac{\sum_j \alpha_j E(\xi_R)}{(\sum_j \alpha_j + J)E(\xi_R)} = \frac{\sum_j \alpha_j}{(\sum_j \alpha_j + J)}
$$

(23)

The denominator in this expression just gives the ratio of all spending to non-wasteful CT spending, $\frac{E(S_R)}{E(NW_{CT})} = \sum_j \alpha_j + J$. We observe $\alpha_{CT} = \frac{E(W_{R,CT})}{E(NW_{R,CT})} = 12.8\%$. We can recover the numerator by solving for $J$ as a function of $\beta_2$ and solving for $\sum_j \alpha_j$ as a function of $J$ and $\beta_1$. Doing so implies that 68.1 \% of medical spending in the economy is wasteful.

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


Table 1: Summary Statistics

<table>
<thead>
<tr>
<th></th>
<th>A. Untested patients</th>
<th>B. Patients with negative tests</th>
<th>C. Patients with positive tests</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>77.6</td>
<td>76.8</td>
<td>76.7</td>
</tr>
<tr>
<td>Female</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>White</td>
<td>0.86</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>Black</td>
<td>0.10</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td>History of PE</td>
<td>0.003</td>
<td>0.006</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Doctor, hospital, and region characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician experience</td>
<td>16.6</td>
<td>16.0</td>
<td>16.6</td>
</tr>
<tr>
<td></td>
<td>(9.0)</td>
<td>(9.0)</td>
<td>(9.0)</td>
</tr>
<tr>
<td>HRR avg spending (in $)</td>
<td>57,040</td>
<td>56,880</td>
<td>56,041</td>
</tr>
<tr>
<td></td>
<td>(7380)</td>
<td>(7220)</td>
<td>(7070)</td>
</tr>
<tr>
<td>Academic hospital</td>
<td>0.3</td>
<td>0.35</td>
<td>0.36</td>
</tr>
<tr>
<td>Top 50 research med. school</td>
<td>0.31</td>
<td>0.3</td>
<td>0.31</td>
</tr>
<tr>
<td>Top 50 primary med. school</td>
<td>0.29</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>No. of observations</td>
<td>6,119,406</td>
<td>133,878</td>
<td>8,604</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doctor experience (in years)</td>
<td>0.00041**</td>
<td>0.00041**</td>
<td>0.00043**</td>
<td>0.00043**</td>
<td>0.00039**</td>
<td>0.00038**</td>
</tr>
<tr>
<td></td>
<td>(0.00007)</td>
<td>(0.00008)</td>
<td>(0.00007)</td>
<td>(0.00007)</td>
<td>(0.00008)</td>
<td>(0.00008)</td>
</tr>
<tr>
<td>Log(avg HRR spend. per benef.)</td>
<td>-0.05339**</td>
<td>-0.03809**</td>
<td>-0.05604**</td>
<td>-0.04931**</td>
<td>-0.05091**</td>
<td>-0.04861</td>
</tr>
<tr>
<td></td>
<td>(0.00843)</td>
<td>(0.01149)</td>
<td>(0.00798)</td>
<td>(0.01138)</td>
<td>(0.00817)</td>
<td>(0.01210)</td>
</tr>
<tr>
<td>Academic hospital</td>
<td>0.00201</td>
<td>0.00060</td>
<td>-0.00014</td>
<td>-0.00111</td>
<td>0.00049</td>
<td>-0.00037</td>
</tr>
<tr>
<td></td>
<td>(0.00159)</td>
<td>(0.00160)</td>
<td>(0.00149)</td>
<td>(0.00152)</td>
<td>(0.00174)</td>
<td>(0.00179)</td>
</tr>
<tr>
<td>Top 50 research med school</td>
<td>0.00142</td>
<td>0.00314</td>
<td>0.00151</td>
<td>0.00289</td>
<td>0.00146</td>
<td>-0.00271</td>
</tr>
<tr>
<td></td>
<td>(0.00197)</td>
<td>(0.00199)</td>
<td>(0.00185)</td>
<td>(0.00185)</td>
<td>(0.00211)</td>
<td>(0.00214)</td>
</tr>
<tr>
<td>Top 50 primary care med. school</td>
<td>0.00094</td>
<td>-0.00076</td>
<td>0.00144</td>
<td>-0.00011</td>
<td>0.00173</td>
<td>0.00051</td>
</tr>
<tr>
<td></td>
<td>(0.00214)</td>
<td>(0.00215)</td>
<td>(0.00203)</td>
<td>(0.00203)</td>
<td>(0.00227)</td>
<td>(0.00226)</td>
</tr>
<tr>
<td>Female doctor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.00157</td>
<td>-0.00093</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00178)</td>
<td>(0.00178)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Controls for comorbidities</th>
<th>No</th>
<th>No</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region fixed effects</td>
<td>None</td>
<td>State</td>
<td>None</td>
<td>State</td>
<td>None</td>
<td>State</td>
</tr>
<tr>
<td>Physician sample</td>
<td>All doctors</td>
<td>All doctors</td>
<td>All doctors</td>
<td>All doctors</td>
<td>EM doctors</td>
<td>EM doctors</td>
</tr>
<tr>
<td>No. of observations</td>
<td>210,750</td>
<td>210,750</td>
<td>210,750</td>
<td>210,750</td>
<td>115,176</td>
<td>115,176</td>
</tr>
<tr>
<td>No. of doctors</td>
<td>32,921</td>
<td>32,921</td>
<td>32,921</td>
<td>32,921</td>
<td>24,273</td>
<td>24,273</td>
</tr>
</tbody>
</table>

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 specific risk factors. Columns 5 and 6 restrict to patients who are evaluated by a physician who specializes in emergency medicine.
Table 3: Calibrating the model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Definition</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s$</td>
<td>0.083</td>
<td>test sensitivity</td>
<td>Lesler et al., 2009</td>
</tr>
<tr>
<td>$fp$</td>
<td>0.05</td>
<td>false positive</td>
<td>Stein et al., 2006</td>
</tr>
<tr>
<td>$MB_{it}$</td>
<td>0.025VSL</td>
<td>medical benefit of testing</td>
<td>Lesler et al., 2009</td>
</tr>
<tr>
<td>$MC_{it}$</td>
<td>0.017VSL</td>
<td>medical cost of testing</td>
<td>Lesler et al., 2009</td>
</tr>
<tr>
<td>$c_{it}$</td>
<td>$300$</td>
<td>financial cost of testing</td>
<td>estimated from Medicare claims</td>
</tr>
<tr>
<td>VSL</td>
<td>1,500,000*</td>
<td>value of a statistical life</td>
<td>Murphy &amp; Topel, 2006</td>
</tr>
</tbody>
</table>

*We allow VSL to vary with age according to the schedule in Murphy & Topel (2006). It is $1.5 \text{ million}$ for a 75 year-old, and declines by approximately $100,000 per year.
Table 4: Estimates of the testing equation for the structural model

<table>
<thead>
<tr>
<th>Independent variables: Patient characteristics</th>
<th>Dependent variable: Chest CT test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 70-74</td>
<td>(1) -0.0003 (2) -0.0008</td>
</tr>
<tr>
<td>Age 75-79</td>
<td>(1) -0.0022 (2) -0.0029**</td>
</tr>
<tr>
<td>Age 80-84</td>
<td>(1) -0.0050 (2) -0.0047**</td>
</tr>
<tr>
<td>Age 85-89</td>
<td>(1) -0.0045 (2) -0.0068**</td>
</tr>
<tr>
<td>Age 90-94</td>
<td>(1) -0.0095 (2) -0.0119**</td>
</tr>
<tr>
<td>Age 95-99</td>
<td>(1) -0.0195 (2) -0.0165**</td>
</tr>
<tr>
<td>Black</td>
<td>(1) -0.0151 (2) -0.0130**</td>
</tr>
<tr>
<td>Hispanic</td>
<td>(1) -0.0111 (2) -0.0082**</td>
</tr>
<tr>
<td>Asian</td>
<td>(1) -0.0030 (2) -0.0003</td>
</tr>
<tr>
<td>Native American</td>
<td>(1) -0.0023 (2) 0.0003</td>
</tr>
<tr>
<td>Other race</td>
<td>(1) -0.0069 (2) -0.0061**</td>
</tr>
<tr>
<td>Unknown race</td>
<td>(1) -0.0044 (2) -0.0023</td>
</tr>
<tr>
<td>Female</td>
<td>(1) 0.0029 (2) 0.0032**</td>
</tr>
<tr>
<td>History of pulmonary embolism</td>
<td>(1) 0.0205 (2) 0.0283**</td>
</tr>
<tr>
<td>History of thoracic aortic dissection</td>
<td>(1) 0.0106 (2) 0.0114**</td>
</tr>
<tr>
<td>History of other aortic dissection</td>
<td>(1) 0.0081 (2) 0.0122**</td>
</tr>
<tr>
<td>History of deep vein thrombosis</td>
<td>(1) 0.0055 (2) 0.0035**</td>
</tr>
<tr>
<td>Previously admitted within 30 days</td>
<td>(1) 0.0044 (2) 0.0036**</td>
</tr>
<tr>
<td>Previously admitted within 7 days</td>
<td>(1) 0.0107 (2) 0.0113**</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doctor experience (in years)</td>
<td>8.79**</td>
<td>7.28</td>
<td>7.99**</td>
</tr>
<tr>
<td></td>
<td>(3.53)</td>
<td>(3.85)</td>
<td>(3.53)</td>
</tr>
<tr>
<td>Log(avg HRR spend. per benef.)</td>
<td>-1746**</td>
<td>-1447**</td>
<td>-1731**</td>
</tr>
<tr>
<td></td>
<td>(243)</td>
<td>(263)</td>
<td>(243)</td>
</tr>
<tr>
<td>Academic hospital</td>
<td>60</td>
<td>61</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>(58)</td>
<td>(63)</td>
<td>(58)</td>
</tr>
<tr>
<td>Top 50 research med school</td>
<td>96</td>
<td>29</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>(85)</td>
<td>(89)</td>
<td>(85)</td>
</tr>
<tr>
<td>Top 50 primary care med. school</td>
<td>58</td>
<td>-27</td>
<td>-58</td>
</tr>
<tr>
<td></td>
<td>(87)</td>
<td>(92)</td>
<td>(87)</td>
</tr>
<tr>
<td>Adjust for pneumonia diagnoses?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Physician sample</td>
<td>All doctors</td>
<td>EM doctors</td>
<td>All doctors</td>
</tr>
</tbody>
</table>

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 accounts 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

<table>
<thead>
<tr>
<th>Welfare metric</th>
<th>Actual testing behavior (1)</th>
<th>Simulated behavior with no over-testing (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent of patients tested with chest CT</td>
<td>0.0364</td>
<td>0.0308</td>
</tr>
<tr>
<td>Number of patients tested with chest CT</td>
<td>73079</td>
<td>62039</td>
</tr>
<tr>
<td>Total costs of testing (millions)</td>
<td>36.1</td>
<td>32.0</td>
</tr>
<tr>
<td>Total benefits of testing (millions)</td>
<td>55.2</td>
<td>70.2</td>
</tr>
<tr>
<td>Net benefits of testing (millions)</td>
<td>19.1</td>
<td>38.2</td>
</tr>
<tr>
<td>Costs per test</td>
<td>494.5</td>
<td>437.9</td>
</tr>
<tr>
<td>Benefits per test</td>
<td>756</td>
<td>1131.1</td>
</tr>
<tr>
<td>Net benefits per test</td>
<td>261.6</td>
<td>615.3</td>
</tr>
</tbody>
</table>

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.
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, $\tau$  

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.
Notes: This figure plots a histogram of the values of $\tau$, 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 Regression of Regional Spending on Wasteful and Non-wasteful CT Spending

The error term in this regression is given by:

\[
\epsilon_R = \xi_R(\sum_j \alpha_j + J) + v_R \sum_j \alpha_j + \sum_j u_{R,j} - \beta_1 \alpha_{CT} \xi_R - \beta_1 \alpha_{CT} v_R - \beta_2 \xi_R - \beta_2 u_{R,j} - \beta_0
\]  

(24)

The coefficients are identified by:

\[
E(\epsilon_R) = E(\xi_R)(\sum_j \alpha_j + J) = \beta_0
\]

\[
Cov(W_{R,CT}; \epsilon_R) = \alpha_{CT}(\sum_j \alpha_j + J)Var(\xi) + \alpha_{CT} \sum_j \alpha_j Var(v_{R,CT})
\]

\[
- \beta_1 \alpha_{CT}^2 Var(\xi) - \beta_1 \alpha_{CT}^2 Var(v_{R,CT}) - \beta_2 \alpha_{CT} Var(\xi)
\]

\[
= \alpha_{CT} Var(\xi) \left[ \sum_j \alpha_j + J - \beta_1 \alpha_{CT} - \beta_2 \right]
\]

\[
+ \alpha_{CT} \left( \sum_j \alpha_j - \beta_1 \alpha_{CT} \right) Var(v_{R,CT}) = 0
\]

\[
Cov(NW_{R,CT}; \epsilon_R) = (\sum_j \alpha_j + J)Var(\xi) - \beta_1 \alpha_{CT} Var(\xi) - \beta_2 Var(\xi) + (1 - \beta_2)Var(u_{R,CT})
\]

\[
= Var(\xi) \left[ \sum_j \alpha_j + J - \beta_1 \alpha_{CT} - \beta_2 \right] + (1 - \beta_2)Var(u_{R,CT}) = 0
\]

(25)

We can further simplify this by using the facts that \( \rho \equiv Cov(NW_{R,CT}, W_{R,CT}) = \alpha_{CT} Var(\xi) \), \( Var(u_{R,CT}) = Var(NW_{R,CT}) - Var(\xi_R) = Var(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \), and

\[
Var(v_{R,CT}) = \frac{1}{\alpha_{CT}} (Var(W_{R,CT}) - \alpha_{CT}^2 Var(\xi_R)) = \frac{1}{\alpha_{CT}} (Var(W_{R,CT}) - \alpha_{CT} \rho).
\]

Plugging these into the above equations yields:

\[
Cov(W_{R,CT}; \epsilon_R) = \rho \left[ \sum_j \alpha_j + J - \beta_1 \alpha_{CT} - \beta_2 \right]
\]

\[
+ \alpha_{CT} \left( \sum_j \alpha_j - \beta_1 \alpha_{CT} \right) \frac{1}{\alpha_{CT}} (Var(W_{R,CT}) - \alpha_{CT} \rho) = 0
\]

\[
Cov(NW_{R,CT}; \epsilon_R) = \frac{\rho}{\alpha_{CT}} \left[ \sum_j \alpha_j + J - \beta_1 \alpha_{CT} - \beta_2 \right] + (1 - \beta_2) \left( Var(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right) = 0
\]
Computing $\Delta = \text{Cov}(W_{R,CT}, \epsilon_R) - \alpha_{CT} \text{Cov}(NW_{R,CT}, \epsilon_R)$ gives:

$$\Delta = \alpha_{CT} \left( \sum_j \alpha_j - \beta_1 \alpha_{CT} \right) \frac{1}{\alpha_{CT}^2} \left( \text{Var}(W_{R,CT}) - \alpha_{CT} \rho \right) - \alpha_{CT} (1 - \beta_2) \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right) = 0$$

We can rearrange this to yield:

$$\beta_1 (\text{Var}(W_{R,CT}) - \alpha_{CT} \rho) = \frac{1}{\alpha_{CT}} \sum_j \alpha_j (\text{Var}(W_{R,CT}) - \alpha_{CT} \rho) - \alpha_{CT} (1 - \beta_2) \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right)$$

which gives:

$$\beta_1 = \frac{1}{\alpha_{CT}} \sum_j \alpha_j (\text{Var}(W_{R,CT}) - \alpha_{CT} \rho) - \alpha_{CT} (1 - \beta_2) \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right)$$

(26)

Plugging this expression back into the expression for $\text{Cov}(NW_{R,CT}, \epsilon_R)$ yields:

$$\frac{\rho}{\alpha_{CT}} \left[ \sum_j \alpha_j + J - \sum_j \alpha_j (\text{Var}(W_{R,CT}) - \alpha_{CT} \rho) - \alpha_{CT}^2 (1 - \beta_2) \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right) \right] - \beta_2$$

$$+ (1 - \beta_2) \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right) = 0$$

(27)

Factoring out $\beta_2$ gives:

$$\beta_2 \left[ \text{Var}(NW_{R,CT}) + \frac{\rho \alpha_{CT} \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right)}{\text{Var}(W_{R,CT}) - \alpha_{CT} \rho} \right] =$$

$$\frac{\rho}{\alpha_{CT}} \left[ \sum_j \alpha_j + J - \sum_j \alpha_j (\text{Var}(W_{R,CT}) - \alpha_{CT} \rho) - \alpha_{CT}^2 \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right) \right]$$

$$+ \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right)$$

(28)

The LHS of this expression simplifies to:

$$\beta_2 \frac{\text{Var}(NW_{R,CT}) \text{Var}(W_{R,CT}) - \rho^2}{\text{Var}(W_{R,CT}) - \rho \alpha_{CT}}$$

(29)
The RHS simplifies to:

\[
\frac{\rho}{\alpha_{CT}} \left[ J + \frac{\alpha_{CT}^2 \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right)}{(\text{Var}(W_{R,CT}) - \alpha_{CT} \rho)} \right] + \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right) + \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right) \] 

\[
= \frac{J \rho}{\alpha_{CT}} + \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right) \left[ 1 + \frac{\rho \alpha_{CT}}{(\text{Var}(W_{R,CT}) - \alpha_{CT} \rho)} \right] 
\]

\[
= \frac{J \rho}{\alpha_{CT}} + \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right) \left[ \frac{\text{Var}(W_{R,CT})}{\text{Var}(W_{R,CT}) - \alpha_{CT} \rho} \right] 
\] 

(30)

Thus, we have:

\[
\beta_2 = \frac{J \rho(\text{Var}(W_{R,CT}) - \rho \alpha_{CT})}{\alpha_{CT}(\text{Var}(NW_{R,CT})\text{Var}(W_{R,CT}) - \rho^2)} + \left( \text{Var}(NW_{R,CT}) - \frac{1}{\alpha_{CT}} \rho \right) \left[ \frac{\text{Var}(W_{R,CT})}{\text{Var}(NW_{R,CT})\text{Var}(W_{R,CT}) - \rho^2} \right] 
\]

\[
= 1 + \frac{\rho^2 - \frac{\rho}{\alpha_{CT}} \text{Var}(W_{R,CT})}{\text{Var}(NW_{R,CT})\text{Var}(W_{R,CT}) - \rho^2} + \frac{J(\frac{\rho}{\alpha_{CT}} \text{Var}(W_{R,CT}) - \rho^2)}{(\text{Var}(NW_{R,CT})\text{Var}(W_{R,CT}) - \rho^2)} 
\]

\[
= 1 + \frac{(J - 1)(\frac{\rho}{\alpha_{CT}} \text{Var}(W_{R,CT}) - \rho^2)}{(\text{Var}(NW_{R,CT})\text{Var}(W_{R,CT}) - \rho^2)} 
\] 

(31)

This implies:

\[
1 - \beta_2 = -\frac{(J - 1)(\frac{\rho}{\alpha_{CT}} \text{Var}(W_{R,CT}) - \rho^2)}{(\text{Var}(NW_{R,CT})\text{Var}(W_{R,CT}) - \rho^2)} 
\] 

(32)

Plugging back in to the expression for \(\beta_1\) gives:

\[
\beta_1 = \frac{1}{\alpha_{CT}} \sum_j \alpha_j (\text{Var}(W_{R,CT}) - \alpha_{CT} \rho) + \alpha_{CT} \frac{(J - 1)(\frac{\rho}{\alpha_{CT}} \text{Var}(W_{R,CT}) - \rho^2)}{(\text{Var}(NW_{R,CT})\text{Var}(W_{R,CT}) - \rho^2)} \]

\[
= \frac{\sum_j \alpha_j}{\alpha_{CT}} + \frac{\rho(J - 1)(\text{Var}(NW_{R,CT}) - \frac{\rho}{\alpha_{CT}})}{(\text{Var}(NW_{R,CT})\text{Var}(W_{R,CT}) - \rho^2)} 
\] 

(33)

Note that \(\rho\) and \(\alpha_{CT}\) are observable, so from the coefficients we can solve for both \(J\) and \(\sum_j \alpha_j\), which allow us to compute the proportion of spending which is wasteful.

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:
\[ \sum_k q^k_{id} NU_k - c_{id} \geq \tau_d \]  

(34)

where \( q^k_{id} \) is the probability of a positive test for condition \( k \) and is given by:

\[ q^k_{id} = x^k_{id} \beta + \alpha^k_d + \eta^k_{id} \]  

(35)

Define \( \theta_d = \sum_k NU_k \alpha^k_d - \tau_d \). Plugging our specifications for the probability of a positive test into the testing equation yields:

\[ \sum_k NU_k x^k_{id} \beta + \theta_d - c_{id} + \sum_k NU_k \eta^k_{id} \geq 0 \]  

(36)

As above define: \( I_{id} \equiv \sum_k NU_k x^k_{id} \beta + \theta_d - c_{id} \). From equation 36, we can compute the expected benefits conditional on testing, which are given by:

\[ E(B_{id} | T_{id} = 1) = \tau_d + I_{id} + g(I_{id}) \]  

(37)

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

Let \( Z^k_{id} \) be an indicator for whether a test for condition \( k \) is positive or negative. If doctors have rational expectations, we must have \( E(q^k_{id} | T_{id} = 1) = E(Z^k_{id} | T_{id} = 1) \). Given these rational expectations and equation 1, we can write the expected benefits as \( E(B_{id} | T_{id} = 1) = \sum_k NU_k E(Z^k_{id} | T_{id} = 1) - c_{id} \). Plugging this into equation 5 and rearranging yields:

\[ \sum_k NU_k E(Z^k_{id} | T_{id} = 1) = \tau_d + c_{id} + I_{id} + g(I_{id}) \]  

(38)

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