NOTE:
The responsibility for the content of this article rests with the author and does not necessarily represent the views of the Institute of Medicine (IOM) or its committees and convening bodies.
There is a long tradition of examining the relationship between healthcare spending and outcomes, and interpreting this association as informative for whether healthcare spending can be increased or decreased without harming patient health. Simple economics reveals that even with perfect risk and price adjustment, this is an unavailing enterprise: comparisons across providers or regions do not tell us the effect of greater spending, but instead confute the joint effect of lower productivity and higher spending: providers who are less productive (achieve worse health outcomes for any level of spending) may attain comparable patient outcomes to other providers by spending more to compensate for their lack of expertise. Moreover comparing spending and outcomes across providers does not measure larger inefficiencies that exist within every provider. In the second part of the paper, I present a prototypical economic model that can distinguish overuse and underuse of treatments as explanations for variations, from the role of expertise in their provision. The model invokes strong assumptions and requires rich data, but underscores the kind of analysis that would be required for the variations literature to inform the spending debate. Our results, developed in the context of heart-attack treatments in the mid-1990s, suggest that expertise varies considerably across hospitals, but a substantial amount of variation in treatment and treatment effectiveness is due to overuse, with the most aggressive hospitals overusing treatment.
I. Introduction

Over 70 years ago, Sir Alison Glower presented variation in tonsillectomy rates to the Royal College of Physicians (Glower, 1938). Since that time, legions of physicians, health-services researchers, statisticians, and economists have asked whether variation in healthcare spending and resource use across regions and providers is associated with improved patient outcomes (see Skinner, 2011 for a review of this literature). Despite the efforts of many to answer this tantalizingly simple question, it has not yielded simple answers; the literature is as contentious as the day when Glower presented his findings.

Studies of variations must grapple with two issues: measurement and frameworks. Much attention has gone into the improving the measurement component, the key dimensions of which are risk-adjustment, price adjustment, and establishing a meaningful unit of geography (for example, patients aren’t constrained by state boundaries, so measurement of variations at the state level is dominated by measurement at the level of hospital referral regions). Substantial energy has gone into understanding whether differences in risk-adjustment or patient populations (Medicare or Commercial; heart-attack patients or hip-fracture patients) are responsible for the different associations between outcomes and spending. Price-adjustment and risk-adjustment that acknowledges the incentives for upcoding are the new frontiers of the variations literature (Gottlieb et. al, 2010 and Song et. al in NEJM). Others believe that the findings are sensitive to the unit of geography used for the analysis—states, regions, or hospitals—and great effort has been poured into studying the relationship at different levels of geographic aggregation.

Here, I make the case that fine-tuning measurement, through superior risk and price adjustment and improved measures of geography, is not enough. Rather, the implicit frameworks used by the literature, both by advocates of ‘more is better’ and ‘more isn’t better,’ are unavailing and can generate inappropriate policy conclusions. The key omitted variable is the technology and expertise with which healthcare inputs are converted into patient outcomes. Expertise in healthcare is the process by which observed medical inputs such as physician services, hospital care, imaging technologies, are combined with other inputs that are almost never measured: advance-planning protocols, decision-aids, safety-culture, diligence, leadership, governance and professionalism. Different delivery systems and HRRs vary in expertise, and consequently, the ability to transform measured inputs like hospital days, hospital spending, imaging and specialists’ visits into outputs. The variations literature has overstated the benefits to medical spending if measured inputs and expertise are positively correlated (as they will be if they are complements), and understated the value of spending if spending more (i.e. using more hospitals days or physician visits) is a substitute for low expertise. Section 2 provides a discussion of these issues, and constitutes the first part of this paper. At the end of this section, I provide an overview of policy takeaways, including the observation that there is similar dispersion in performance in other industries, including concrete, where there aren’t third-party payers, government reimbursement systems, or asymmetric information between buyers and sellers. Whatever the source of productivity dispersion in healthcare, we should be wary of developing solutions which are unique to healthcare.
In Section 3, I develop a prototypical economic model of treatment choice with hospitals varying in their expertise that can make the distinctions that the variations literature is trying to make. The model offers a way to distinguish overuse and underuse—which imply that we should spend less and spend more respectively—from expertise, which is the use of superior production methods. The model makes strong assumptions and requires excellent data, but underscores the kind of analysis that is required to inform policy. It illustrates the substantial gap between where the variation literature is, and where it needs to be before its findings rise to the level of being policy relevant. The model requires an excellent ability to risk-adjust and price-adjust, so efforts to improve these dimensions of measurement are still valuable.

The key idea for separating underuse from overuse is to examine the benefit from treatment for last patient who received the treatment (the marginal patient). In contrast to overuse and underuse, higher expertise would result in all patients benefiting more from care, and so it is rational for such a system to perform more treatment in less appropriate patients. Consequently, the marginal patient will differ across providers based on their level of expertise. Providers should deliver treatment to the point where its marginal value is zero; not doing more when the marginal patient has positive benefit, implies underuse. Analogously, if the marginal patient is harmed by care (the treatment effect is negative), then providers should do less. The key empirical challenge is to develop a tractable statistical framework for identifying the marginal patient and making inferences about this patient. We present this framework and estimate its key parameters using data from heart-attack treatments. Our results, which may be unique to heart-attack treatments in the era of thrombolytic treatments, suggest that expertise varies considerably across hospitals, but a substantial amount of variation in treatment and treatment effectiveness in due to overuse.

2. Frameworks for Interpreting Variations

The variations literature is fundamentally focused on efficiency—can we deliver better health for less spending? If the answer to this question is yes, the system is inefficient, but it is helpful to understand why, and make a distinction between productive inefficiency and allocative inefficiency. Productive inefficiency occurs when a patient’s outcomes could be improved with a given level of inputs; for the same inputs, some providers get worse outcomes than other providers. In this paper, I will refer to this form of efficiency as expertise, but it should be noted that expertise is a broad notion and encompasses everything from greater diligence and superior triage, to the use of electronic medical records and hand-washing. Allocative inefficiency is when a healthcare system either withholds beneficial treatment from some patients (underuse), or provides low-benefit or even unwittingly harmful treatment to other patients (overuse). For example, if we observe a much higher rate of cardiovascular surgery in Region A compared to Region B, we might suspect allocative inefficiency, whether arising from too much care among inappropriate populations in Region A, or the rationing of care in Region B. But it could also be that healthcare providers in Region A do more because they are productively efficient, getting better results from a given procedure, or choosing more appropriate patients.
Which form of inefficiency is responsible for the variation across providers? Both forms of inefficiency are possible, and both are possible within the same delivery system. The variations literature has never been precise about which form of inefficiency it is identifying, and therefore conflates the effects of spending more or less with more (or less) expertise or productive efficiency. This omission is important, because the policy prescriptions for reducing productive inefficiencies in healthcare (same as lowering variation in expertise) may be very different than that which reduces overuse and underuse. For example, the general belief in “practice guidelines” implicitly assumes that a uniform practice standard would improve allocative efficiency by reducing underuse and overuse. But this policy would not help if practice variation were due to differences in expertise. With differences in expertise, it is actually sensible for some providers to do more than other providers, because they are better at providing that care. Similarly, tying payments to outcomes sounds wonderful, but may encourage overuse in highly efficient delivery systems, whose superior outcomes would exist despite overuse.¹

Figure 1 illustrates the conventional thinking. Assume that the data is perfectly risk and price adjusted so that these considerations do not affect this discussion. The figure relates healthcare spending (inputs) to outputs such as patient survival or quality-adjusted-life-years. Figure 1 illustrates the association between spending on factor inputs (health resources such as physicians, scans, and hospitals) on the horizontal axis and survival/quality of life on the vertical axis. A concave health production function illustrates the aggregate health that is achievable for a given level of inputs for a particular delivery system. The production possibility frontier assumes diminishing returns to spending: Higher levels of spending are associated with inputs that generate increasingly smaller health benefits; computed tomography (CT) scanners, proton-beam accelerators, and chemotherapy for metastatic cancers will have lower benefit per dollar spent relative to aspirin for heart-attack patients or antiretroviral therapy for HIV and AIDS patients. Under diminishing returns, physicians are first giving treatments to the patients who benefit most from the treatment, and then moving on to patients with lower benefit. Technological advances can push the production possibility frontier up and out. The frontier falls after a point because ‘too much’ healthcare can actually harm patients because of iatrogenic injury.

The variations literature is focused on whether different regions, hospitals, or delivery systems, are described by points A, B, C, and D. System A does too little (underuse) and system D does too much (overuse). In system A, increasing spending would yield better outcomes for patients. In contrast, lives and money could be saved by making regions D do less. A finding of no relationship between spending and outcomes (as in between B and C, or A and D) is interpreted as ‘flat of the curve medicine.’ In the case of comparing A to D one might be tempted to think that we should make system D like system A, but that misses the point that A needs to do more and D less because of allocative inefficiency. The conventional test for underuse is to see whether spending and outcomes are positively correlated and for overuse to see whether spending and outcomes are negatively correlated.

¹ Some policies would reduce both forms of inefficiency: for example, a move towards bundled payments, or greater competition in the healthcare industry.
Figure 2 illustrates how the above framework arrives at the wrong policy recommendation when there are expertise differences in how inputs are converted to outputs. In both panels of this figure, each system has its own technology or expertise for producing care. Some systems are more superior than others in that they are able to deliver better outcomes for the same amount of spending because of differences in expertise. In Panel A, note that every system does ‘too little’ and that outcomes would be universally improved by every system ‘doing more’ in the form of using more inputs such as hospital days, imaging and physician services. In other words, each system underutilizes intensive care. Despite this underuse, the negative association between spending and outcomes (i.e. the dashed line) falsely suggests that high cost systems have overuse. In Panel B, the opposite conclusion is illustrated. Here, every system is doing too much and each should do less, but the association between spending and outcomes is positive which misleads us into thinking that underuse is a concern.

Figure 1 is the implicit framework for learning the effect of spending on health. This view is shared by advocates of wasteful spending in healthcare (e.g. Fisher et al, (2003a,b) and Fuchs, 2004)), but also by researchers who have disagreed with his findings (Ong et.al 2009; Bach, 2010; Silber et.al, 2010, and Romley et.al, 2011). By comparing outcomes and spending, each of these papers assume that they’re comparing regions on the same production function. Indeed, had Fisher’s detractors found the same results as in Fisher et.al they would have concluded the same thing as Fisher—that more spending isn’t automatically beneficial. But Figure 2 offers a cautionary tale for why the simple approach, while highly intuitive, is misleading for public policy. Panel 2A illustrates the key concern with the Fisher studies. In these, more intensive care is not associated with better outcomes, but the reality is that outcomes could be improved in reality by doing more. Simply cutting intensive spending at the high-cost (but low productivity) provider, without fixing the underlying productivity problem, will result in worse patient outcomes. Panel 2B highlights the concern that ‘more is better’ papers have to grapple with—it may be that more spending is associated with superior outcomes, but as the exhibit illustrates, each system is harming patients because of overuse and every system should be doing less. And even if one doesn’t believe the possibility of overuse, the figure illustrates how productivity improvements allow systems to do better even by not spending more.

The implications of different regions or systems having different production functions may be even more important than what has been discussed in Figure 2. In Figure 3, I illustrate a pathological case that is particularly problematic for authors who have argued that more spending is good for patient health. Here, Hospital Y has adopted technologies in the wrong order—that is, it has adopted less productive technologies first, and more productive technologies next. This pattern would show up in any hospital (or system) that adopted proton-beam therapy or unproven cancer therapies before fully adopting beta-blockers or rescue angioplasty for heart-attack treatments. This could result from insurers paying for these intensive technologies more generously than other more therapeutic but sometimes revenue-shrinking interventions such as care management, reduced errors, or lower admissions. In this scenario, even though Hospital Y has lower average productivity and higher costs than Hospital X, the last dollars spent for Hospital Y actually yield more value than the last dollars spent for Hospital X. In this situation, more spending is actually better and reducing spending
will harm patients in Hospital Y, but that conclusion misses the tremendous opportunities from improving productivity and doing more for less.

Evidence

Skinner, Staiger and Fisher (Health affairs 2006) and Chandra and Staiger (JPE 2007) were the first to demonstrate that different regions were on different production functions. In Chandra and Staiger (2007), this happens because delivery systems specialize in how they deliver care. Some specialize in technologically intensive treatments, whereas others may specialize in less expensive treatments. Because of this specialization, even though two systems or countries may have similar health outcomes (with one achieved at substantially lower cost), simply cutting spending in the high-cost system will not allow it to achieve the results of a low-cost system, but it will almost certainly harm patient health in the process. Finding that two systems have similar outcomes at dissimilar spending levels is therefore not informative about whether we should spend more or less on health care. But this question cannot be answered in the presence of productive heterogeneity; a first step must be to understand why the two systems have different production functions and whether and at what cost the production function might be changed. The variations literature does not answer these questions.

Another example comes from my work with Mary Beth Landrum and others (Landrum et. al, 2008). We examine the care of patients with colorectal cancer, and find that high-spending regions are more likely than other regions to use recommended care but are also more likely to use discretionary and non-recommended care, the latter of which has adverse outcomes for patients. This is another way of saying that different regions use different production functions (i.e. Figure 2 is a better approximation that Figure 1). At least for this type of cancer, finding that spending more is cost effective might result in our spending more on colorectal cancer when the appropriate policy response would be the opposite. The same can be said if we find would that spending is ‘cost-ineffective’--the majority of the spending may have been effective but diluted by the presence of an expensive but potentially harmful treatments. Absent knowing what is being purchased with the extra money, it is difficult to use aggregate productivity calculations to ascertain whether we should be spending more or less.

Another stream of evidence that illustrates the role of productivity differences comes from overtime studies of healthcare spending of the type resented in Cutler, Rosen and Vijan, 2006, as well as Murphy and Topel, 2006. This research compares outcomes and spending over time (as opposed to across regions) and attributes better outcomes to higher spending. Chandra and Skinner (2012) offer a dissenting view of this interpretation, arguing that much of the improvement in outcomes comes from ‘high average productivity treatments’ while treatments of dubious clinical value are responsible for the costs. Our idea is illustrated with an example: suppose that there are two treatments A and B. Treatment A is useful and costs a dollar, but treatment B is

---

2 Cutler (2004) writes, “Money matters in health care as it does in few other industries. Where we have spent a lot, we have received a lot in return.”
useless and costs $100,000. For various reasons, the use of treatment B may be correlated with treatment A or other unobserved factors contributing to good health over time or across regions. In this case, an aggregate productivity calculation that regresses survival on total spending would find that the extra spending ‘was worth it’. Such conclusions leave policy makers thinking that all is right with the world, and that more spending is efficient, and perhaps even necessary. But in reality we could have slashed almost 100 percent of spending without harming patients.

What is the evidence for this concern? Ford et al. (2007) notes that 44 percent of the reduction in coronary deaths between 1980 and 2000 were the consequence of changing risk factors related to behaviors rather than health care per se. The implication of this is that aggregate productivity studies may overstate the overall improvement in life expectancy attributable to health care expenditures given that behavioral factors accounted for nearly half of the survival improvement. In terms of what caused the gains in coronary deaths: 35 percent of the decline in mortality was the consequence of inexpensive but highly effective treatments such as aspirin, beta blockers, blood-thinning drugs, anti-hypertensives, diuretics, and pharmaceuticals such as ACE inhibitors, anti-cholesterol drugs (statins), and thrombolytics (“clot-busters”). The marginal cost of these inputs is modest. So almost 80 percent (44 percent + 35 percent) of the survival was caused by relatively inexpensive treatments. Innovations such as angioplasty (stents), bypass surgery, cardiac rehabilitation and cardio-pulmonary resuscitation (such as automated defibrillators) explained less than 12 percent of the mortality decline, but are responsible for an enormous portion of the costs. And so, regressing outcomes on spending might find that the extra spending was worth it, but that would lose sight of the fact that virtually all the survival gains came from ‘home run’ technologies such as beta-blockers and aspirin.

2.2 What do we learn from studies of geographic variation?

In light of this discussion, we should be skeptical about drawing any policy implications about spending more or less from the variations literature. But its finding are still valuable, for regardless of whether the association between spending and outcomes was found to be positive, negative or zero, the papers in this literature demonstrate tremendous variability in performance at the same level of spending. To illustrate, see Figure 4 (see color insert), which shows the relationship between hospital-level spending and one-year survival for patients who were diagnosed with heart attacks, hip fracture, or colorectal cancer (Chandra et al. 2010). Regardless of the overall relationship between survival and outcomes (zero in this exhibit), there is large variation in the performance of hospitals around the average relationship (shown by the horizontal line in the figure), highlighting the potential for reducing both productive and

---

3 One gauge of whether we have picked up the effect of spending more (that is providing more treatment) from the effect of expertise is to ask whether estimates of more medical spending can be reconciled with results from clinical trials. If these estimates exceed those from trials, it is likely that we have confounded expertise with spending. Similarly, if the estimates are lower than those from trials, it is possible that we are picking up the effect of overuse in addition to expertise.
allocative inefficiencies in the delivery system. This dispersion would exist even if the cloud relating outcomes to inputs were sloped upwards or downwards. Similar variability in performance are noted by Ong et.al (2009)—where at the same level of hospital inputs there is wide variation in patient mortality.

The fact that different papers find different results is more likely due to both productive inefficiency (doing better with the same amount) and allocative inefficiency (doing too much or too little). One can interpret this literature as providing evidence of the need and opportunity to improve the productivity of the delivery system. Unfortunately, this literature has been unable to separate the relative roles of these two forms of efficiency, and absent this distinction one cannot say whether spending should be increased or decreased through fee cuts, capacity constraints, or geographically based reimbursements systems.

One question that naturally follows this discussion is whether the variation literature correctly identifies the entirety of productive and allocative efficiencies. I believe that it does not because it ignores all the inefficiencies that occur inside every delivery system, state, or HRR. In other words, this literature ignores productive inefficiency that is present everywhere—and that is the bulk of the inefficiency. To the extent that providers everywhere struggle with clinical uncertainty, decision fatigue, a lack of information, and malpractice pressures, the geographic variations literature will miss these inefficiencies. In Chandra, Cutler and Zong (2011), we explore the case for explanations that may be pervasive across all healthcare systems. Situational factors and behavioral biases such as the availability heuristic or status-quo biases, may be at work everywhere, and are likely to be a much smaller part of regional variations in care. In diagnosis, the availability heuristic suggests that a physician who has just seen a patient with influenza may be more likely to make the diagnosis of influenza for the next patient who walks through the door with a cough, even if this latter patient has a rare lung disease. The reverse may be true if there was a recent case of lung cancer. This status quo bias may be further driven by a self-reinforcing confirmation bias in which prior successes from one treatment induce the provider to continue using that treatment. Confirmation bias describes the idea that people interpret new information in light of their pre-existing biases (Myers and Lamm, 1976 and Halpern, 1987). Specialization may be a source of this confirmation bias (Sommers et al., 2008b). A specialist that has experienced success with a particular treatment—for example, radical prostatectomy for patients with localized prostate cancer—will judge that treatment to be effective in a wider range of patients, likely even wider than objective literature would indicate. Whether this judgment is substantiated by any positive spillovers in the learning-by-doing sense, however, is an open question. These situational factors are evident in conceivably every decision to prescribe or receive health care, and understanding them may offer the greatest potential improvements for the productivity of health care spending. But their presence is not identified by the variations literature.

Finally, one must recognize that the conventional wisdom in health-policy seems to be that the driving forces behind large average productivity differences are various idiosyncratic institutional features of the healthcare sector that effectively reduce competitive pressures on providers. In particular, the combination of uninformed consumers who lack knowledge of the quality differences across providers, generous
health insurance that insulates consumers from the direct financial consequences of their healthcare consumption decisions, and public sector reimbursement that provides little incentive for productive efficiency by providers, is widely believed to dull the basic disciplining force of demand-side competition that exists in most other sectors. Echoing and summarizing this view, Cutler (2010) notes:

“There are two fundamental barriers to organizational innovation in healthcare. The first is the lack of good information on quality. Within a market, it is difficult to tell which providers are high quality and which are low quality… Difficulty measuring quality also makes expansion of high-quality firms more difficult… The second barrier is the stagnant compensation system of public insurance plans.”

In a similar vein, Skinner (2011) states in his *Handbook of Health Economics* article on regional variations in healthcare that low productivity providers are:

“…unlikely to be shaken out by normal competitive forces, given the patchwork of providers, consumers and third-party payers each of which faces inadequate incentives to improve quality or lower costs…”

As a result of this thinking it is tempting for health services researchers to propose solutions for productivity improvements that are unique to healthcare (e.g. the need to adopt electronic medical records). This “healthcare exceptionalism” view stands in marked contrast to a very large empirical literature outside of the healthcare sector that has documented extensively – almost without exception – enormous differences in average productivity across producers within narrowly defined industries (see Bartelsman and Doms 2000, Syverson 2011 and references therein). For example, on average within narrow US manufacturing (4-digit SIC) industries, the average productivity ratio between the 90th and 10th percentile producer is about 2 to 1 (Syverson 2004a). This literature has shown how various features of the distribution of productivity across firms can be used to make inferences about the existence and extent of market competition.⁴

In Chandra, Syverson, Finkelstein and Scarny (2012) we examine the extent of competition in healthcare markets, specifically the market for hospital treatment of Medicare heart attacks. Specifically, we calculate market-level statistics about productivity differences and allocation across different producers. We find that productivity dispersion across hospitals is similar to what we see in the market for ready-made concrete, which is a product that is far more homogeneous than healthcare, but is also geographically differentiated (suppliers in Louisville KY do not supply concrete to purchasers in Lexington KY). In healthcare, hospitals with lower productivity have smaller market share and are more likely to close. Qualitatively, our results indicate that competitive market forces do affect resource allocation in the hospital sector. Quantitatively, across a variety of different metrics our results suggest

⁴ The theoretical basis for this notion can be found, for example, in Ericson and Pakes (1995) and Melitz (2003); examples of empirical work using this approach include Olley and Pakes (1996), Disney, Haskel, and Heden (2003) and De Loecker (2011).
that the impact of competition on resource allocation, and the extent of misallocation across producers, are similar in the hospital sector to what has been found in U.S. manufacturing.

We view our results as suggesting that the healthcare sector may not be as idiosyncratic as the conventional wisdom appears to hold. In this sense, our results are in the same spirit as work by Skinner and Staiger (2007, 2009), who find that areas of the country that were innovative as early adopters of hybrid corn in the 1930s and 1940s were also early adopters of beta blockers for treating heart attacks at the beginning of the current century, suggesting a role for a common factors across sectors of “innovativeness” within an area. More generally, our reading of the results is that they offer some cautious optimism that the broader productivity research agenda on the factors that determine firm-level productivity—which has both produced useful insights and continues to be an area of active on-going research (Syverson 2011)—may be useful for guiding and informing research on improving productivity in the healthcare sector. We consider further studies of the determinants of productivity in the healthcare sector—guided by insights learned from the broader productivity literature—to be an important and fruitful direction for further work. Much like the firms in manufacturing, it is entirely possible that the keys to improving productivity in healthcare lie in management practices, and the leadership and professionalism of providers. This would suggest a very different set of policy levers than targeting spending.

To summarize these findings for public policy, I believe that the variations literature is fundamentally uninformative about whether we should spend more or less. On the one hand, it may have drastically overstated the benefits of intensity (by conflating intensity or spending with expertise). On the other, if high spending providers are doing more to make up for a lack of expertise, the variations literature has understated the amount of productive inefficiency. It also has no way of measuring inefficiencies that exists inside every delivery system, which are probably first-order because of the extent to which system-wide factors such as malpractice and Medicare reimbursement affect all systems. By missing this dimension of inefficiency, my sense is that this literature has generally understated inefficiency. At the risk of suggesting a slightly glib metaphor, the variations literature is like a thermometer or stethoscope—it can highlight areas where the system is struggling, and identify areas of medicine where providers disagree on the right thing to do. But such measurement does diagnose why it is struggling (clinical uncertainty, poor triage, inadequate decision support, weak leadership) and what to do about it. Absent a diagnoses for whether overuse or underuse or productive inefficiencies plague our system, a focus on the right amount of spending is premature. Finally, given that productivity dispersion in healthcare is similar to productivity dispersion in other sectors, it would be wise to stay away from designing policies that are unique to healthcare.

5 Because of the robust correlations between spending and capacity, it is popular to assert that doing more is ‘supply sensitive,’ as in caused by the presence of doctors and beds. This is possible, but far from established. Far more likely, capacity and utilization are both outcomes of the same production function.
3 Distinguishing Overuse, Underuse and Expertise

In this section, I present a simple empirical model that can distinguish between alternative explanations for variations in treatment rates, based on a behavioral model in which hospitals choose to treat patients if the benefit from treatment exceeds a hospital-specific threshold. This work is based on unpublished research with Douglas Staiger.

We develop a model that acknowledges patient heterogeneity in the benefit for medical treatments; indeed this is key to our test. In our model, differences in the use of a treatment across hospitals may be due to greater benefits of treatment in some hospitals (expertise), withholding of beneficial treatment in some hospitals (underuse), or providing harmful treatment in other hospitals (overuse). Evidence of expertise differences is evidence of productive inefficiency in healthcare, whereas evidence of overuse and underuse is evidence of allocative inefficiency. Expertise, underuse, and overuse are identified based on differences across hospitals in both their treatment rates and the treatment effect on patient survival.

Using detailed chart-data on heart attack treatments (which illustrates the data-intensive needs of this approach), we find that expertise varies considerably across hospitals, but a substantial amount of variation in treatment is due to overuse. In this work model, we equate the use of reperfusion with spending more. This was done for analytical convenience (we were able to verify that we can replicate the causal effect of reperfusion on survival) and can certainly be relaxed. We do not claim that our results can be generalized to outside our setting, but offer this approach as an example of what needs to be done to learn more about components of productive inefficiency in healthcare spending. While we apply this framework to heart-attack treatments, it is quite general and could be applied to separate overuse, underuse and expertise in other healthcare settings such as cancer treatments.

3.1 Intuition

Our intuition is based on Gary Becker’s original framing for how to test for discrimination in labor-markets. The two panels of Figure 5 offer the central intuition. In the first panel, I explain the intuition for testing for overuse and underuse: we are comparing outcomes for patients in two different hospitals (or regions). Providers rank patients in order of diminishing appropriateness and work down the distribution of patients (this is a testable assumption and is verified below). They should stop delivering treatment when the marginal benefit is zero. If care is underutilized in one system compared to the other, outcomes for the marginal patients receiving care in the less-intensive region should be higher than outcomes for similar patients who are receiving care in the more intensive region.\(^6\) This is because the less intensive region only performed the treatment in its most appropriate patients, and passed up on providing valuable care to

---

\(^6\) Statisticians and epidemiologists will recognize that ‘average benefits for patients who received the treatment’ is the same as the ‘treatment on the treated effect,’ in models where there is treatment effect heterogeneity. It is different from the average treatment effect, but the two will be the same if there is no heterogeneity in benefit across patients.
patients who would still benefit. In contrast, if care is over-provided in one system (that is, the system is operating in the range of negative marginal benefits), then the system that over-treats will have lower average benefits for patients who receive the treatment.

In contrast to overtreatment and under treatment as explanations for geographic variation in healthcare, there is also heterogeneity in expertise. This is illustrated in the second panel of Figure 5, where the entire marginal benefit curve is shifted out is more productive delivery systems. Because of this improved productivity, it makes sense for these delivery systems to do more in less appropriate patients. In the Figure, I have illustrated expertise as an entire shift in the marginal benefit curve, but in practice it is possible for expertise to only benefit some patients—e.g. older patients or cancer patients.

To identify overuse and underuse, we compare outcomes for two patients who have the same estimated propensity to receive treatment. This propensity to receive treatment may depend on where the patient received treatment—possibly because some providers are better at delivering care; expertise should increase appropriateness. If there is expertise in the production of healthcare, a comparison of identical patients (patients with the same propensity) who received treatment in two regions should reveal higher benefits in the region that did more. One can identify overuse by looking at the average benefit care in less appropriate patients. This is where overuse would happen, and if there are lower benefits to less-appropriate patients in high-use areas, then we have identified overuse.

3.2 Economic Model

We assume that treatment is provided to each patient whenever the expected benefit from the treatment exceeds a minimal threshold. Thus, in the terminology of Heckman, Urzua and Vytlacil (2006), our model allows for essential heterogeneity where the decision to provide treatment to each patient is made with knowledge of their idiosyncratic response to treatment. Within this framework, there are two ways in which a patient’s hospital could affect treatment choice. First, the expected benefit of treatment for a given patient may vary across hospitals, reflecting each hospital’s idiosyncratic level of expertise in providing the treatment. Second, the minimum threshold for receiving care may vary across hospitals, as determined by local incentives and norms at each hospital. From the patient’s point of view, treatment should be provided whenever the expected benefit from treatment exceeds zero. Therefore, there is underuse of the treatment in hospitals that set a minimum benefit threshold above zero, and overuse in hospitals that set a minimum threshold below zero.

Let $B_{ih}$ represent the expected benefit from treatment for patient $i$ at hospital $h$. Benefit is the gain or improvement in survival relative to not receiving a treatment. We focus on the health benefits of the treatment, which would include any reduction in mortality or morbidity that was expected from the treatment, e.g. the impact of the treatment on Quality Adjusted Live Years (QALYs). In principal, the benefit could also incorporate the expected impact of the treatment on the patient’s medical cost, and capture the health benefits net of costs. However, in our application we focus on survival for simplicity.
Suppose that the expected benefit from treatment depends on the hospital’s expertise ($\alpha_h$), observable patient characteristics ($X_{ih}$) such as age, medical history, and lab results, and other factors that are known to the medical care provider when making the treatment decision but unobserved by the econometrician ($\epsilon_{ih}$):

$$B_{ih} = \alpha_h + X_{ih} \beta + \epsilon_{ih}$$

Each patient receives treatment if the expected benefit from treatment exceeds a minimal threshold ($\tau_h$), where the threshold varies across hospitals due to local incentives or norms. This corresponds to a simple Roy model of treatment allocation, where a patient receives treatment if the gain from the treatment exceeds a threshold. Being treated at a more expert hospital raises the benefit of the treatment to the patient, and in principle we can allow the effect of expertise to vary with patient characteristics. Assuming that $B_{ih}$ captures the total net benefit to the patient of providing treatment, then the optimal decision from the patient’s perspective would let $\tau_h = 0$ and provide treatment whenever the benefits to the patient exceed zero. There is underuse if $\tau_h > 0$, since patients with positive benefits are under the threshold and do not receive treatment. There is overuse if $\tau_h < 0$, since patients with negative benefits (who would do better without treatment) are above the threshold and receive treatment.

This decision process implies a very simple structure that determines both the probability of treatment as well as the expected benefit conditional on being treated (the treatment-on-the-treated parameter). The probability of receiving treatment is just the probability that expected benefits exceed the minimum threshold:

$$\Pr(Treatment_{ih} = 1) = \Pr(B_{ih} > \tau_h) = \Pr(-\epsilon_{ih} < I_{ih}),$$

where $I_{ih} = X_{ih} \beta + (\alpha_h - \tau_h)$

Equation (2) highlights that differences in treatment rates across hospitals, holding all else equal, can be due to either differences in hospital expertise ($\alpha_h$) or differences in the treatment threshold ($\tau_h$) reflecting over or underuse. A hospital may be more likely to provide treatment because of greater expected benefits of treatment ($\alpha_h > 0$) or because of a lower benefit threshold for providing care ($\tau_h < 0$). Conversely, even if treatment rates were the same across hospitals, there could still be overuse or underuse if, say, hospitals with greater expected benefits of treatment set a correspondingly higher threshold for providing care ($\alpha_h = \tau_h > 0$). Thus, because variation in treatment rates across hospitals masks variation in hospital expertise and hospital treatment thresholds, such variation cannot by itself say anything about overuse or underuse. In other words, while the conventional wisdom is to interpret variation in hospital treatment rates or spending as saying something about differences in their treatment threshold (as in flat of the curve models of provider behavior), this inference ignores the role of expertise in also explaining that variation.
However, overuse and underuse can be identified separately from hospital expertise if information on the treatment effect among the treated population is available. The treatment-on-the-treated parameter is defined as:

\[
E\left(B_{th} \mid \text{Treatment }_i = 1\right) = E\left(B_{th} \mid -\varepsilon_{ih} < I_{ih}\right) = X_{ih}\beta + \alpha_h + E\left(\varepsilon_{ih} \mid -\varepsilon_{ih} < I_{ih}\right)
\]

(3)

Noting that \(X_{ih}\beta + \alpha_h = I_{ih} + \tau_h\), we can rewrite Equation (3) as:

\[
E\left(B_{th} \mid \text{Treatment }_i = 1\right) = \tau_h + g(I_{ih}),
\]

(4)

where \(g(I_{ih}) = I_{ih} + E\left(\varepsilon_{ih} \mid -\varepsilon_{ih} < I_{ih}\right)\)

Equation (4) states that in the absence of any difference across hospitals in the minimum threshold to receive care, two patients receiving treatment who have the same propensity to get the treatment (same \(I\)) will have the same expected benefit from the treatment. This relies on the assumption that \(g(I)\) depends only on the index, i.e. we must assume a single-index selectivity model, so that the truncated mean of the error in equation (4) depends only on the truncation point \((I)\). This would not be the case if the distribution of the unobservable factors determining treatment \((\varepsilon)\) differed across hospitals, which in turn would happen if some hospitals were better at measuring these unobservable factors. Examining equation (4) makes it clear that hospitals with larger unobservables will have larger treatment on the treated effects (which will look like having a higher treatment threshold) because of the higher values of conditional error term.

Identificaton Conditional on the Estimated Propensity

If the propensity to get the treatment (or equivalently the index \(I\)) can be estimated directly from Equation (2), we can identify differences in the minimum threshold from an estimate of the treatment-on-the-treated parameter: For patients with the same propensity, those treated at a hospital with a higher minimum threshold (a higher \(\tau_h\)) will have a larger treatment effect. Moreover, because \(g(I_{ih})\) goes to zero as the propensity to receive treatment goes to zero (or equivalently as \(I_{ih}\) gets increasingly negative), we can also identify overuse \((\tau_h < 0)\) or underuse \((\tau_h > 0)\): There is overuse when the treatment effect for the lowest propensity patients is negative, and underuse when the treatment effect for the lowest propensity patients remains positive.

It is important to note that our model does not imply that the treatment-on-the-treated parameter is the same for all patients treated in hospitals with the same minimum threshold. In fact, Equation (4) implies that the treatment-on-the-treated effect will tend to be larger among patients that have a higher propensity to be treated (since \(g(I_{ih})\) is increasing in \(I_{ih}\)). The treatment effect is the same only for patients with the same propensity to be treated and treated in hospitals with the same minimum threshold.

Since hospital expertise \((\alpha_h)\) raises the propensity to be treated, it will also raise the treatment effect. The key difference, however, is that differences in hospital expertise
have an impact on treatment effects by shifting the propensity to be treated (and, therefore, \(g(I_{ih})\)), while differences in the minimum threshold have an impact on treatment effects that is independent of the propensity.

**Graphical Intuition**

The graphical intuition for our model can be seen in Figure 6. The expected benefit from treatment (B) is given on the vertical axis, while the propensity of being treated (which depends on I) is given on the horizontal axis. The top curve in Figure 1 represents the treatment-on-the-treated effect for a patient with a given propensity that is treated in a hospital with a high minimum threshold for treatment \(\tau_{\text{high}} > 0\), i.e. it represents \(E(B_{ih} \mid B_{ih} > \tau_h) = \tau_h + g(I_{ih})\). The lower curve represents the same thing for a hospital with a low minimum threshold \(\tau_{\text{low}} < 0\). Treatment-on-the-treated approaches the minimum threshold (\(\tau_{\text{high}}\) or \(\tau_{\text{low}}\)) for a patient with a low propensity of being treated (a very negative I), since no patient is ever treated with a benefit below this threshold. For a patient with a high propensity of being treated (a very positive I), truncation becomes irrelevant and the treatment-on-the-treated effect asymptotically approaches the unconditional benefit of treatment. However, conditional on a patient’s propensity, the treatment effect is always higher in the hospital with the higher threshold. In this figure, differences in hospital expertise would show up as a movement along the curves – changing the propensity of patients to be treated (and therefore the treatment-on-the-treated effect), but not affecting treatment effects conditional on propensity.

### 3.3. Heart-Attacks

**Biology and Treatments**

Heart attacks (more precisely, acute myocardial infarction (AMI)) occur when the heart-muscle (the myocardium) does not receive sufficient oxygen, because of a blockage in one of the coronary arteries which supply blood to the heart. We focus our empirical work on the treatment of AMI for a number of reasons, recognizing throughout that the focus on heart-attacks may not extend to the rest of healthcare:

First, cardiovascular disease, of which heart attacks are the primary manifestation, is the leading cause of death in the US. A perusal of the leading medical journals would indicate that heart attack treatments are constantly being refined, and a large body of trial evidence points to significant therapeutic gains from many of these treatments. In this context, variation in treatments across hospitals may directly translate into lost lives, and there is a rich tradition of studying variation across hospitals in treatments and outcomes after heart attacks.

Second, as a consequence of what is known about heart attack treatments from randomized controlled trials, and more specifically for our setting, the benefits from reperfusion, we are able to assess whether our regression estimates of the benefits from reperfusion are comparable to those found in the medical literature, or whether they are
confounded by selection-bias. We focus on reperfusion, where our use of chart data allows us to replicate the RCT evidence that is summarized by the Fibrinolytic Therapy Trialists' Collaborative Group (1994). Chart data provides comprehensive documentation on the patient’s condition at the time that the treatment decision is made, and therefore minimizes the possibility that unobserved clinical factors related to a patient’s survival are correlated with treatment.

Third, because mortality post-AMI is high (mortality rates at 30 days are nearly 20 percent), a well-defined endpoint is available to measure the productivity of hospitals. This would not be true if we focused on treatment variation for more chronic conditions such as diabetes, chronic obstructive pulmonary disease, or arthritis.

Our fourth reason for focusing on heart attacks is that it is an acute condition for which virtually all patients are hospitalized at a nearby hospital and receive some medical care. Moreover, during the acute phase of the heart attack the therapeutic emphasis is on maximizing survival, which is achieved by timely reperfusion, and hospital staff (not patients and their families) make treatment decisions. The fact that patients are generally taken to the nearest hospital for immediate treatment, makes it likely that a patient’s choice of hospital is exogenous and not driven by differences across hospitals in expertise or the treatment threshold. This feature of heart attack treatments would not be true, for example, in the case of cancer therapies, where two clinically identical patients may chose different providers based on their evaluation of the provider’s expertise and standards for providing treatment.

Data

Because acute myocardial infarction is both common and serious, it has been the topic of intense scientific and clinical interest. One effort to incorporate evidence-based practice guidelines into the care of heart attack patients, begun in 1992, is the Health Care Financing Administration's Health Care Quality Improvement Initiative Cooperative Cardiovascular Project (CCP). Information about more than 200,000 patients admitted to hospitals for treatment of heart attacks in 1994/1995 was obtained from clinical records. The CCP is considerably superior to administrative/claims data (of the type used by McClellan et al. (1994)) as it collects chart data on the patients—detailed information is provided on laboratory tests, enzyme levels, the location of the myocardial infarction, and the condition of the patient at the time of admission. Detailed clinical data were abstracted from each patient’s chart using a standard protocol. Further details about the CCP data are available in Marciniak et al. (1998), and O’Connor et al. (1999). The choice of sample and variables is identical to what we used and described in Barnato et al. (2005) and Chandra and Staiger (2007, 2010). Following the clinical literature, we define a patient to have received reperfusion if any of these therapies was provided within 12 hours of the heart attack. In our data from the mid-1990s, over 90 percent of patients receiving reperfusion received thrombolytics.

Our estimation strategy relies on using the richness of the CCP data to invoke a ‘selection on observables’ assumption to estimate treatment on the treated for reperfusion therapy. We feel comfortable making this assumption because of our ability to use CCP data to match the estimates from clinical trials. A summary of nine trials
was published in the journal *Lancet* by the Fibrinolytic Therapy Trialists' Collaborative Group (FTTCG, 1994). This was the same time-period as the CCP data and each trial evaluated fibrinolytic therapy in heart-attack patients. Across these nine trials, reperfusion within 12 hours reduced 35-day mortality by 2.3 percentage points (s.e.=0.5 percentage points) among patients over age 65 relative to a base mortality rate of 18.5% in the control group. With the CCP, we estimated the effect of reperfusion (pooling across all demographic groups) on 30-day survival as 2.3 percentage points (s.e.=0.3 percentage points) relative to an average mortality rate of 18.8% in our sample. We take this evidence as supporting the case for simple regression models that invoke a ‘selection on observables’ assumption.

### 3.4 Estimation

**The Propensity to Receive Treatment**

We use data on heart attack treatments to estimate the key components of our model, using receipt of reperfusion within 12 hours of the initial heart attack as our treatment. The propensity to receive treatment (\( I \) in the theoretical model) is estimated from a random effect logit model that regresses whether a patient received reperfusion within 12 hours of the heart attack on all the CCP risk-adjusters listed in the appendix (\( X_{ih} \)) and a random hospital effect (\( \theta_h \)) that is assumed to be normally distributed:

\[
\Pr(\text{reperfusion } n_{ih} = 1) = F(X_{ih}\beta + \theta_h), \quad \text{where} \quad \theta_h \sim N(0, \sigma^2_\theta)
\]

(5)

Where the function \( F(.) \) represents the logistic function. Equation 5 is the empirical analog of Equation 2 in our model, where the hospital random effect is the difference between expertise and the threshold (\( \theta_h = (\alpha_h - \tau_h) \)) as in Equation 2 Estimating equation 5 (using xtmelogit in Stata 11) yields maximum likelihood estimates of the coefficients \( \hat{\beta} \) and the standard deviation of the hospital random effect \( \hat{\sigma}_\theta \), and posterior estimates of the hospital random effects \( \hat{\theta}_h \). We combine these to form an estimate of the propensity index for each patient \( \hat{I}_{ih} = X_{ih}\hat{\beta} + \hat{\theta}_h \).

**Estimation of Treatment Effects Conditional on the Estimated Propensity**

We focus on patient survival as the outcome of interest which captures the key benefits of treatment. Survival is measured as a binary variable that measures survival to a certain date (e.g. survival to 30 days). Our model and Equation 4 in particular, suggests

---

7 This estimate used propensity score weighting to make the distribution of propensities for the untreated resemble that of the treated (that is, weighting the untreated by \( \Pr(\text{Reperfusion})/1-\Pr(\text{Reperfusion}) \)).
that the effect of reperfusion on survival should be heterogeneous across patients and hospitals. This suggests estimating models of the following form (where $F(.)$ represents the logistic function):

$$\Pr\left(Survival_{ih} = 1\right) = F\left(\text{reperfusion } n_{ih} \delta_{ih} + X_{ih} \gamma + \varphi_{ih}\right), \text{ where } \delta_{ih} = \tau_{h} + g(I_{ih})$$

(6)

In equation 6, survival for all patients depends on patient risk adjusters ($X_{ih}$) and a hospital effect that captures general skill of the hospital staff ($\varphi_{ih}$). The coefficient on reperfusion ($\delta_{ih}$) is the empirical analog of Equation 4 in our model, and captures the survival benefit of reperfusion. Our model says the effect of reperfusion on survival depends only on a patient’s propensity index ($I_{ih}$) and on the minimum treatment threshold at the hospital ($\tau_{h}$).

Equation 6 suggests a simple test for whether the minimum threshold for treating a patient ($\tau_{h}$) varies across hospitals. Conditional on a patient’s propensity index ($I_{ih}$), variation across hospitals in the benefit of reperfusion is due to differences in the minimum threshold. This minimum threshold should be negatively correlated with the hospital random effect $\theta_{h}$ from the propensity equation (since $\theta_{h} = (\alpha_{h} - \tau_{h})$). Intuitively, according to our model, hospitals with higher minimum thresholds for treatment should both treat fewer patients (conditional on $X$) and have higher benefits to treatment (conditional on $I$).

To implement this test, we use estimates of $I_{ih}$ and $\theta_{h}$ from the propensity equation (5), and approximate the treatment effect with a linear function $\delta_{ih} = \lambda_{0} + \lambda_{1} I_{ih} + \lambda_{2} \theta_{h}$, yielding an estimating equation:

$$\Pr\left(Survival_{ih} = 1\right) = F\left(\text{reperfusion } n_{ih} \lambda_{0} + \left(\text{reperfusion } n_{ih} \lambda_{1} I_{ih}\right) \lambda_{1} + \left(\text{reperfusion } n_{ih} \lambda_{2} \theta_{h}\right) \lambda_{2} + X_{ih} \gamma + \varphi_{ih}\right)$$

(7)

If hospitals vary in their expertise ($\alpha_{h}$) but not in their minimum threshold for treating a patient ($\tau_{h}$), then our model implies $\lambda_{2} = 0$; controlling for patient propensity, the treatment effect is unrelated to the hospital effect in the propensity equation ($\hat{\theta}_{h}$). Alternatively, if hospitals vary in their minimum threshold but not in expertise, then we expect $\lambda_{2} < 0$; controlling for patient propensity, the treatment effect is smaller in hospitals with a high propensity to treat. If both expertise and the minimum threshold vary across hospitals, then we expect $\lambda_{2} < 0$ unless there is a strong positive association between expertise and the minimum threshold (since the sign of $\lambda_{2}$ depends on $\text{cov}(\tau, \theta) = \text{cov}(\tau, \alpha - \tau) = \text{cov}(\tau, \alpha) - \text{var}(\tau)$).

Equation 7 provides two other tests of hospital behavior. First, if hospitals choose patients for treatment based on the benefit of the treatment, then the treatment effect
should be increasing in the patient’s propensity index ($\lambda_i > 0$). Second, the treatment effect among patients with a low propensity index is an upper bound estimate of each hospital’s minimum threshold, since $\delta_{ih} = \tau_h + g(I_{ih})$ and $g(.)$ approaches zero as the propensity index falls. Therefore, if the treatment effect among low-propensity patients is negative in some hospitals, this is evidence of overuse ($\tau_h < 0$).

We explore a number of alternative specifications to Equation 7 in order to check the robustness of our results. First, in some specification we include hospital fixed effects (using fixed-effect logit models) in order to capture general differences in hospital skill affecting survival of all patients ($\phi_h$). Second, in some specifications we estimate $g(I)$ non-parametrically with 100 indicator variables for the percentiles of $I$ interacted with reperfusion (and included directly in $X$), thus allowing the return to reperfusion to vary flexibly with a patient’s propensity to receive reperfusion. Finally, we also use semi-parametric methods (described in more detail in a future appendix) to flexibly estimate how the effect of reperfusion on survival varies with $\hat{i}_{ih}$ and with $\hat{\theta}_h$.

3.5 Results

In Table 1 we report some basic characteristics of our sample overall, and by whether the patient received reperfusion within 12 hours of admission to the hospital. In our sample, 19% of patients received reperfusion within 12 hours of admission for a heart attack. Overall, 81% of patients were still alive 30 days after admission, but survival was higher for patients receiving reperfusion (86%) than for patients who did not receive reperfusion (80%). However, much of the difference in survival between these two groups was due to differences in underlying health and pre-existing conditions, rather than the result of reperfusion. Patients receiving reperfusion were younger, and much less likely to have pre-existing conditions such as congestive heart failure, hypertension, diabetes, and dementia.

The coefficients on the patient-level variables are consistent with the medical literature, with reperfusion being less likely among patients with pre-existing conditions and who are older (based on full age-sex-race interactions not reported in the table), and also depending on the location and severity of the heart attack. The estimated standard deviation of the hospital effect is 0.44 (Std. Err. = 0.01), which implies that a one standard deviation in the hospital effect increases the logodds of receiving reperfusion by 0.44, which would increase an average patient’s probability of receiving reperfusion from 19% to 26%. Thus, there is sizable variation across hospitals in the rate at which they provide reperfusion to observationally similar patients. The model is able to predict much of the hospital-level variation, with the posterior prediction of each hospital’s effect on reperfusion having a standard deviation of 0.30 in our data.

Estimates of Treatment Effects Conditional on the Estimated Propensity
In Table 2 we present results from estimating the treatment effect of reperfusion conditional on the estimated propensity using Equation 7:

\[
\Pr(Survival_{ia} = 1) = F\left(\text{reperfusion } n_{ia} \lambda_0 + \text{reperfusion } n_{ia} \hat{I}_{ih} \lambda_1 + \text{reperfusion } n_{ia} \hat{\theta}_h \lambda_2 + X_{ia} \gamma + \psi_{ia}\right)
\]

Recall from the earlier discussion that the key parameters of interest are those associated with reperfusion \((\hat{\lambda}_0, \hat{\lambda}_1, \hat{\lambda}_2)\). Table 2 reports estimates of these parameters, which together determine the treatment effect: \(\delta_{ih} = \lambda_0 + \hat{\lambda}_1 \hat{I}_{ih} + \lambda_2 \hat{\theta}_h\). The top panel reports coefficients from various specifications using simple logit models (clustering standard errors at the hospital level), while the bottom panel reports coefficients from the same specifications adding hospital fixed effects (conditional logit models).

As a baseline, the first column of Table 2 reports estimates that do not condition on the patient’s propensity, but allow the effect of reperfusion to interact with the estimated hospital effect from the propensity equation \((\hat{\theta}_h)\). Since this effect is mean zero, the coefficient on reperfusion gives the treatment effect for an average hospital, and is 0.229 and highly significant in the model without hospital fixed effects. The coefficient on the interaction with the hospital effect from the propensity equation is negative, but not statistically significant. The results with hospital fixed effects are very similar. Since these specifications do not condition on the propensity, the coefficient on the interaction with the hospital effect (which is part of the propensity) is biased in the positive direction, and not a strong test of whether hospitals differ in their minimum treatment threshold.

The second column of Table 2 allows the effect of reperfusion to interact with the patient’s propensity to receive reperfusion (the index, \(\hat{I}_{ih}\)). To help with interpretation, we have normalized the index so that a value of 0 refers to the average patient receiving reperfusion. Thus, the coefficient on reperfusion is an estimate of the effect of reperfusion on an average patient receiving reperfusion treated at an average hospital. The coefficient on the interaction of reperfusion with the propensity index is positive and highly significant in specifications with or without hospital fixed effects, implying that the treatment effect of reperfusion on survival is increasing in the patient’s propensity index \((\lambda_1 > 0)\) as predicted by our model. The coefficient on this interaction implies that an increase in the propensity index of one (about one standard deviation of the propensity index in the treated population) is associated with roughly a doubling of the treatment effect. Thus, it appears that hospitals are choosing patients for treatment based on the benefit of the treatment, and the heterogeneity in the treatment effect is large relative to the average treatment effect.

As expected, the coefficient on the interaction of reperfusion with the hospital effect from the propensity equation becomes more negative and statistically significant in the specifications that condition on the propensity index. The coefficient is similar in column 3, where we non-parametrically control for the interaction of reperfusion with a set of 100 dummies for each propensity percentile. In other words, conditional on a patient’s propensity, the treatment effect is smaller in aggressive hospitals with a high
propensity to treat \( (\lambda_0 < 0) \). As discussed earlier, this finding suggests that hospitals vary in their minimum threshold, and more aggressive hospitals with lower minimum thresholds for treatment treat more patients (conditional on \( X \)) and have lower benefits to treatment (conditional on \( I \)). The estimated coefficients suggest that a one standard deviation increase in the hospital effect from the propensity equation (about 0.3) lowers the return to reperfusion by about .06-.09.

As discussed above, the treatment effect for patients with a low propensity index is an (upper bound) estimate of each hospital’s minimum threshold, and can be used to identify overuse \( (T_k < 0) \). Based on the estimates from Table 2, we can use \( \delta_{ih} = \lambda_0 + \lambda_1 \hat{I}_{ih} + \lambda_2 \hat{\theta}_h \) to calculate the expected treatment effect for patients with a low propensity index admitted to a given hospital. For example, consider a “low propensity” patient who is one standard deviation below average in their propensity index (about -1). Using estimates from the parametric model without hospital fixed effects, a low propensity patient at an average hospital would have a treatment effect near zero \((0.345-1*0.299=.006)\). But a low propensity patient at a hospital that was one standard deviation above average (about 0.3) in its hospital effect from the propensity equation would have a negative treatment effect \((0.345-0.299*1-0.258*0.3=-0.077)\). In other words, for many patients in our sample, our estimates imply overuse (a negative treatment effect).

The parametric model estimated in Table 2 imposes linearity, and may not provide accurate predictions of treatment effects for patients with low propensity. Therefore, we explore our results non-parametrically in Figures 7 and 8. In the left-hand panel of Figure 7 we plot the estimated survival benefit from reperfusion (and 95% confidence interval) against the hospital effect from the propensity equation using a locally-weighted logit model to estimate the reperfusion effect (controlling non-parametrically for the propensity index as was done in column 3 of Table 2). The right-hand panel of Figure 7 is the analogous plot but estimated only for low-propensity patients whose propensity index implied that they had a probability of receiving reperfusion of below 20%. Both plots show a clear downward slope, with lower benefit from treatment for patients treated by hospitals with higher random effects in the propensity equation. Among all patients (the left-hand plot), the estimated survival benefit from reperfusion is positive for all hospitals, although it is small and not significant in hospitals with the highest treatment rates (those 2 standard deviations above average, with \( \hat{\theta}_h =0.6 \)). In contrast, among the lowest propensity patients (the right-hand plot), only hospital’s with the lowest treatment rates are estimated to have survival benefits from reperfusion that are near to zero. The estimated survival benefit from reperfusion is negative and significant in hospitals with the highest treatment rates, suggesting that there is overuse in these hospitals and, as a result, we were able to identify substantial subsets of patients who were harmed by reperfusion treatment.

Figure 8 plots the estimated survival benefit from reperfusion against a patient’s treatment propensity index for hospitals in the lowest (left-hand side) and highest (right-hand side) terciles of the estimated hospital effect from the propensity equation \( (\hat{\theta}_h) \).
Both plots show a strong upward slope, with higher benefit from treatment for patients with a higher propensity to receive reperfusion. But at every propensity, the benefits of reperfusion are lower in the top-tercile hospitals. At the lowest propensity levels, the survival benefits from reperfusion are significantly negative for the top-tercile hospitals, suggesting that there is overuse among these hospitals. In the bottom-tercile hospitals, the estimated survival benefits from reperfusion for the lowest propensity patients are less negative and not significantly different from zero, which is consistent with appropriate use of reperfusion in these hospitals.

4. Conclusions

The reasons for variation in treatment rates across hospitals serving similar patients may be due to greater benefits of treatment in some hospitals (expertise), withholding of beneficial treatment in some hospitals (underuse), or providing harmful treatment in other hospitals (overuse). The variations literature does not make this distinction, but the empirical work in this paper distinguishes between these explanations, using a simple behavioral model in which hospitals choose to treat patients if the benefit from treatment exceeds a hospital-specific threshold. We distinguished between expertise, underuse, and overuse based on differences across hospitals in both their reperfusion rates and the effect of reperfusion on patient survival in a sample of heart attack patients. Our results suggest that expertise varies considerably across hospitals, but a substantial amount of variation in treatment and treatment effectiveness in our data was due to overuse. We find that the most aggressive hospitals are also at highest risk of overuse.

We do not mean to suggest that our results apply to all of healthcare, or even to all of heart-attack care. Our goal was simply to point out that it takes a fair bit of economic structure to parse out overuse from underuse and expertise. People may be uncomfortable with this structure, but as the discussion of Sections I and II reminds us, absent structure of this type, the variations literature does not shed light on whether we should spend more or less on healthcare. As noted in Section 2, it may have drastically overstated the benefits of intensity (by conflating intensity or spending with expertise). On the other hand, as the example of reperfusion therapy for heart-attacks patients illustrates, if providers are doing more to make up for a lack of expertise, the variations literature has understated the amount of productive inefficiency. Finally, much like the variations literature, our analysis also missed inefficiencies that exist inside every hospital. We suspect that these may be large.
References


Lanndrum MB, Meara ER, Chandra A, Guadagnoli E, Keating NL. Is Spending More Always Wasteful? The Appropriateness Of Care And Outcomes Among Colorectal Cancer Patients. *Health Affairs* January 2008 vol. 27no. 1 159-168 [http://content.healthaffairs.org/content/27/1/159.abstract](http://content.healthaffairs.org/content/27/1/159.abstract)


Figure 1. The ‘Flat of the Curve’ framework for learning the effect of spending on health.
Figure 2. Alternative Interpretations of the Variations Literature. In Panel 2A more intensive care is not associated with better outcomes, but outcomes could be improved by each system doing more. In Panel 2B more spending is associated with superior outcomes, but each system is harming patients because of overuse and every system should be doing less.
Figure 3. How productive inefficiency could create a ‘more is better’ result
Figure 4. Association between one-year survival and spending at the hospital level for patients with heart attacks, hip fractures, and colorectal cancer—all conditions with limited discretion in diagnosis. We combined these measures into a single quality dimension and a single cost dimension for the 3,804 hospitals in our sample. All spending numbers are reported in 2005 dollars and include both hospital spending and physician spending. Figure taken from Chandra et al. (2010).
Figure 5. Panel A illustrates underuse and overuse. If care is underutilized in one system compared to the other, average outcomes for patients receiving care in the less-intensive region should be higher than average outcomes for patients who are receiving care in the more intensive region. Panel B illustrates expertise, which shifts out the entire marginal benefit curve through more productive delivery systems.
Figure 6: How the Expected Benefit from Treatment Varies with the Propensity to Get Treatment and the Treatment Threshold. The figure illustrates the relationship between the expected benefit from treatment, \( E(B|B > \tau) \), on the vertical axis, and the propensity of being treated on the horizontal axis. The thick curves represent the treatment-on-the-treated effect for a patient with a given propensity. It approaches the minimum threshold (\( \tau \)) for a patient with a low propensity of being treated. The top curve represents a hospital with a high treatment threshold (underuse) and the bottom curve represents a hospital with a low treatment threshold (overuse).
Figure 7: Survival Benefit from Reperfusion According to Hospital Effect on Treatment Propensity, All patients (Panel A) and Low-propensity patients (Panel B). The left-hand panel plots the estimated survival benefit from reperfusion (and 95% confidence interval) against the hospital effect from the propensity equation using a locally-weighted logit model to estimate the reperfusion effect (controlling non-parametrically for the propensity index as was done in column 3 of Table 2). The right-hand panel is the analogous plot estimated only for low-propensity patients whose propensity index implied that they had below a 20% probability of receiving reperfusion.
Figure 8: Survival Benefit from Reperfusion According to Patient’s Treatment Propensity, Low-Treatment-Rate (Panel A) and High-Treatment-Rate (Panel B) Hospitals. The figures plot the estimated survival benefit (and 95% confidence intervals) from reperfusion against a patient’s treatment propensity index for hospitals in the lowest (left-hand side) and highest (right-hand side) terciles of the estimated hospital effect from the propensity equation.
Table 1: Means of Selected Variables, Overall and By Reperfusion

<table>
<thead>
<tr>
<th>Variable</th>
<th>Full Sample</th>
<th>Received Reperfusion w/in 12 hours</th>
<th>No Reperfusion w/in 12 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survival 30 days post-AMI</td>
<td>81%</td>
<td>86%</td>
<td>80%</td>
</tr>
<tr>
<td>Reperfusion w/in 12 hours</td>
<td>19%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Age</td>
<td>77</td>
<td>73</td>
<td>77</td>
</tr>
<tr>
<td>Previous diagnoses:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congestive Heart Failure</td>
<td>22%</td>
<td>7%</td>
<td>25%</td>
</tr>
<tr>
<td>Hypertension</td>
<td>62%</td>
<td>56%</td>
<td>63%</td>
</tr>
<tr>
<td>Diabetes</td>
<td>30%</td>
<td>23%</td>
<td>32%</td>
</tr>
<tr>
<td>Dementia</td>
<td>6%</td>
<td>2%</td>
<td>7%</td>
</tr>
<tr>
<td>Number of observations</td>
<td>138,957</td>
<td>25,876</td>
<td>113,081</td>
</tr>
</tbody>
</table>
Table 2. Logit Estimates of the Effect of Reperfusion on 30-day Survival, Conditional on the Estimated Propensity and the Hospital Effect from the Propensity Equation.

<table>
<thead>
<tr>
<th></th>
<th>Not Conditional on Propensity</th>
<th>Conditional on Propensity (Parametric)</th>
<th>Conditional on Propensity (Non-Parametric)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Without Hospital Fixed Effects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reperfusion &lt;12 hours</td>
<td>0.229 (0.027)</td>
<td>0.345 (0.027)</td>
<td>non-parametric</td>
</tr>
<tr>
<td>* propensity index</td>
<td></td>
<td></td>
<td>(0.018)</td>
</tr>
<tr>
<td>Reperfusion &lt;12 hours</td>
<td>-0.094 (0.078)</td>
<td>-0.258 (0.080)</td>
<td>-0.310 (0.080)</td>
</tr>
<tr>
<td>* Hospital effect from propensity equation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>With Hospital Fixed Effects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reperfusion &lt;12 hours</td>
<td>0.214 (0.026)</td>
<td>0.328 (0.027)</td>
<td>non-parametric</td>
</tr>
<tr>
<td>* propensity index</td>
<td></td>
<td></td>
<td>(0.018)</td>
</tr>
<tr>
<td>Reperfusion &lt;12 hours</td>
<td>-0.052 (0.074)</td>
<td>-0.211 (0.076)</td>
<td>-0.254 (0.077)</td>
</tr>
<tr>
<td>* Hospital effect on treatment propensity</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Dependent variable is the 30 day survival. The top panel reports coefficients from various specifications using simple logit models (clustering standard errors at the hospital level), while the bottom panel reports coefficients from the same specifications with hospital fixed effects (conditional logit models). Standard errors in parentheses. Models include all CCP risk-adjusters. Column 3 includes 100 percentiles of I interacted with the receipt of Reperfusion.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>(Standard Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reperfusion Equation:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Std. Dev. Of Hospital-Level intercept</td>
<td>0.442</td>
<td>(0.013)</td>
</tr>
<tr>
<td><strong>30-day Survival Equation:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reperfusion &lt;12 hours</td>
<td>0.265</td>
<td>(0.026)</td>
</tr>
<tr>
<td>Reperfusion &lt;12 hours  * (Xβ)</td>
<td>0.276</td>
<td>(0.018)</td>
</tr>
<tr>
<td><strong>Hospital-level intercept</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.199</td>
<td>(0.017)</td>
</tr>
<tr>
<td>Correlation with reperfusion intercept</td>
<td>-0.100</td>
<td>(0.073)</td>
</tr>
<tr>
<td><strong>Hospital-level coefficient on reperfusion</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.307</td>
<td>(0.056)</td>
</tr>
<tr>
<td>Correlation with reperfusion intercept</td>
<td>0.035</td>
<td>(0.112)</td>
</tr>
<tr>
<td>Correlation with survival intercept</td>
<td>-0.381</td>
<td>(0.151)</td>
</tr>
<tr>
<td><strong>Transformed Estimates:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital minimum treatment threshold</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.327</td>
<td>(0.055)</td>
</tr>
<tr>
<td>Correlation with survival intercept</td>
<td>-0.321</td>
<td>(0.150)</td>
</tr>
<tr>
<td>Hospital expertise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.451</td>
<td>(0.045)</td>
</tr>
<tr>
<td>Correlation with minimum threshold</td>
<td>0.390</td>
<td>(0.144)</td>
</tr>
<tr>
<td>Correlation with survival intercept</td>
<td>-0.330</td>
<td>(0.115)</td>
</tr>
</tbody>
</table>

Note: See text for discussion of estimation method and equations.