

OUTCOMES ASSESSMENT AND HEALTH CARE REFORM

Amitabh Chandra^{1,3,4}

Elliott Fisher^{1,5,6}

Jonathan Skinner^{1,2,3}

Draft v12: September 1, 2007

A fundamental restructuring of U.S. health care is inevitable, even if we don't know what form the new system will take. But any program providing universal coverage will still need to solve the twin problems of poor quality and rampant growth in expenditures. Some observers have argued in favor of a greater reliance on technology assessment and cost effectiveness, making sure that every dollar spent yields maximal health benefit. We argue that outcomes assessment, or measuring directly the efficiency of a health care system (whether hospital, physician group, or larger network) is instead the ideal approach to solving these twin problems of lagging quality and cost growth. We first contrast technology assessment and outcomes assessment, and then present an example of outcomes assessment using hospital-level risk-adjusted mortality and Medicare expenditures during 1992-2004. In this exercise, focusing on just a few large hospitals, we demonstrate that the standard measures of process or technology assessment, such as use of effective drugs or surgical interventions, do not appear to explain much in the way of growth in actual mortality or expenditures. We argue that new methods to measure outcomes and costs accurately, coupled with the development of "accountable care organizations," hold the greatest promise for improving the efficiency and long-term viability under any health care reform.

¹ Community and Family Medicine, Dartmouth Medical School, Hanover NH; ²Department of Economics, Dartmouth College, ³ National Bureau of Economic Research, Cambridge, MA; ⁴ Kennedy School of Government, Harvard University, Cambridge MA; ⁵ Department of Medicine, Dartmouth Medical School, ⁶ Veteran's Affairs Outcomes Group, White River Junction, VT. We thank without implicating Victor Fuchs and John E. Wennberg for helpful comments, and are grateful to NIA PO1-AG19713 for financial support.

1. Introduction

Everyone agrees that the U.S. health care system is in crisis, with high and rapidly growing health care costs, poor quality, rising numbers of uninsured patients, and an efficiency index lagging that in Albania (Evans, et. al., 2003). There is no lack of proposed solutions to fix health care in the U.S., with reforms ranging from a single-payer national health insurance to increased reliance on Health Saving Accounts and other market-based cures. However, all of these reforms must confront a serious problem: how to control excessive expenditures and restrain cost growth while improving the quality of care. Nor is this a problem that is unique to the United States. All developed countries have been struggling with health care absorbing an ever-larger fraction of government and private budgets (Kotlikoff and Hagist, 2005).

One potential solution is to rely more heavily on technological assessment and principles of cost effectiveness to help government or private providers draw the line on excess spending for procedures that are unlikely to yield much if any benefit (e.g., Pearson, 2007). It is difficult to argue with the principle that each dollar spent in the health care sector should deliver something of real value to the recipients. In practice, however, technology assessment has faced a variety of challenges, ranging from technical -- ensuring that new therapies are compared to the best alternative and that all trials are published (e.g., Hayward et. al., 2005) – to socio-cultural, that voters even agree that rationing on the basis of cost-effectiveness is a good idea (Nord, et. al., 1995). But even if a vigorous and effective technology assessment program were implemented, a key question remains: Is technology assessment sufficient to do the job?

In this paper, we argue that a complementary approach to technology assessment, *outcomes* assessment, or measuring the efficiency of health care systems, is also necessary for the success of any health care reform. With few exceptions, results from randomized trials cannot be generalized to the entire universe of patients who are potential candidates for treatment with new technologies. Patients vary in the amount of benefit that they obtain for a given intervention. Even if they do not, differences in time and risk preferences may cause patients to value the same therapeutic benefit very differently. Similarly, health care providers differ with respect to how well they perform specific procedures, with low-volume or non-academic hospitals often exhibiting worse outcomes than those observed in state-of-the-art clinical trials (Wennberg, et. al., 1998). Over time, providers may learn how to perform a procedure better, or invest in complementary technologies that enhance its benefit, factors that may not be reflected in centralized decisions regarding the procedure.¹ In contrast, outcomes assessment explicitly attempts to capture the overall impact of multi-dimensional treatment strategies, and identifies those health care systems that both adopt appropriate technologies and perform them successfully.

A related concern is that a substantial fraction of health care spending is devoted to services that are not easily brought under traditional approaches to cost effectiveness analysis (CEA) because they are not provided to treat specific abnormalities. These range from variations in the intensity of management of chronic disease to different approaches in diagnosing patients with new symptoms or concerns. The remarkable variations in per-patient spending observed across academic medical centers with similar outcomes are largely due to differences in use of

¹ It is also possible that biases may partially cancel out; a procedure may not be initially cost-effective in community hospitals, but gradually become more effective in those hospitals as physicians gain experience with the procedure.

largely discretionary services such as the frequency of physician office visits or specialist consultation, differences in the relative intensity of imaging services, and how much time similar patients spend in institutional settings (Fisher et al., 2004). There is some evidence that suggests the growth of these services, as opposed to treatments that are administered in an inpatient setting (and amenable to evaluation by CEA), account for the lion's share of cost growth in U.S. healthcare.

To provide a concrete example of outcomes assessment, we use claims data on Medicare patients experiencing a heart attack (acute myocardial infarction, or AMI) during the period 1992-2004. We aggregate individual patient outcomes (as measured by one-year mortality) and costs (one-year expenditures) for several large hospitals to derive hospital-specific measures of efficiency. Thus we can ask how closely standard measures of technological proficiency, for example the use of surgical procedures or β blocker use, is associated with performance measures in terms of both outcomes and costs. We demonstrate that these traditional measures of technology adoption explain little in subsequent outcomes, raising the question of what factors *do* cause some hospitals to become so much more cost-effective.

Given the ability to risk adjust in a reliable way (Krumholz et. al., 2007), and the ease of measuring the primary outcome of survival, our focus on acute myocardial infarction provides an undemanding case study for outcomes assessment. A more pertinent discussion is whether other treatments such as those for multiple chronic ailments or “preference sensitive care” (where patient preferences should affect the choice of treatment), can ever be amenable to reliable outcomes assessment, particularly given the difficulties associated with the precise measurement of functional outcomes and preferences. Nor is it entirely clear how one can translate our focus on outcomes assessment into action, particularly when the future landscape of health care in the U.S. is so uncertain. We argue, however, that establishing local organizational accountability for quality, outcomes and costs -- through the establishment of “accountable care organizations” (Fisher, et. al., 2006), at the hospital or physician group level -- is a critical component for the success of any health care reform initiative that hopes to solve the twin challenges of rising costs and persistent gaps in quality.

2. Contrasting Technology Assessment and Outcomes Assessment

Technology assessment focuses on the evaluation of therapies designed to treat a specified biological perturbation or abnormality. For example, there are many studies comparing specific treatments for the different manifestations of heart disease, such heart attacks or congestive heart failure. Much of the gains in survival following heart attacks can be attributed to the pioneering randomized trials establishing the efficacy of low-cost treatments such as aspirin and beta blockers, while discouraging other treatments with no proven benefits (e.g., Swan-Ganz catheters) or that even caused harm (lidocaine). More generally, Garber (1994), Weinstein (1996), and others have argued that technology assessment – and its implementation through the use of cost-effectiveness analysis – is central to controlling costs and improving quality. Pearson (2007) provides a comprehensive review of both the advantages and shortcomings of technology assessment.

Cost-effectiveness analysis (CEA) is a closely related component of technology assessment, because it provides a framework for ranking alternative technologies in a reasonable way, so that treatments with the highest value per dollar be provided first, and that society should

“work down” the list until they come to the point where the health-related outcome is no longer deemed sufficiently cost-effective to justify payments. In theory, decisions about the cut-off point for cost-effectiveness should be made by “society,” or failing that by institutions that pay the bills such as a national health insurance agency.² In the United Kingdom, the National Institute for Health and Clinical Effectiveness (NICE) has shown success in implementing cost-effectiveness principles particularly in approving or disapproving pharmaceuticals (Pearson, 2007).

	I	II	III	IV	V
N of people with disease X saved	10	20	30	40	50
N of people with disease Y saved	8	6	4	2	0
Total saved	18	26	34	42	50
<i>Percentage of Survey Respondents Choosing Each Option</i>					
	5%	27%	48%	14%	6%

Source: Nord, et. al., 1995, Table 4.

Table 1: Five Different Ways to Allocate \$1 Million Dollars, with Lives Saved of People with Diseases X and Y, and Most Preferred Options as Chosen by Survey Respondents

There does not appear to be any institution in the U.S. that is willing or able to make these difficult decisions. For example, Redberg (2007) documented the willingness of Medicare carriers to pay for Computed Tomography Angiography (CTA), despite both the substantial risk inherent from the high levels of radiation, and the almost complete lack of evidence that it prevents adverse outcomes.

Why hasn't technology assessment been given a more prominent role, especially given the clear concerns about rising health care costs? One intriguing strand of the literature argues that the weak link is the public – that voters, at least Australians ones, simply do not agree with the guiding principle of cost-effectiveness analysis. Results from a survey conducted Down Under is presented in Table 1, where respondents were asked about hypothetical choices between treating people with Disease X, which is treated cheaply, versus Disease Y requiring more expensive treatments (Nord, et. al., 1995). Respondents understood the tradeoff, and that spending a fixed budget to save people with Disease Y would lead to fewer overall lives saved. Five options (I through V) are shown in Table 1, with total lives saved in the third row. Just 6 percent of the population chose the cost-effective solution (V), about as many as choose the least cost-effective approach (5 percent). Nearly half choose III, leading to just 34 lives saved instead of the maximum of 50.³ The respondents viewed the cost-effective approach as unfair because it failed to insure against the risk of contracting a cost-ineffective disease.

The Oregon experiment in cost-effective rationing could be viewed as another example of disconnect between the principles of cost-effectiveness and voter preferences, but there were a variety of factors in this reform that complicates the interpretation of this one natural experiment (Oberlander, 2007). A more important point is that voters in the U.S. rarely need to make the

² Two other approaches are “appropriateness evaluation” and “strength of evidence” which do not rely explicitly on costs but instead on how appropriate the care is (Garber, 1994). Because of difficulties in defining “appropriate” care, we do not pursue these approaches.

³ This result could also reflect a “central tendency” of respondents to choose the median (III) option.

difficult choices described in this survey, given the high prevalence of potential cost saving by scaling back on procedures with no proven benefits.

It seems clear that there are both large potential benefits, along with significant hurdles, that face health policy makers seeking to implement technology assessment in the current U.S. health care system. But even if we were to implement the principles of technology assessment, it would still not solve many fundamental problems regarding costs and quality – or more generally, the efficiency – of the health care system. Ultimately, the value of a technology for a specific patient is tied to (a) characteristics of the patient who gets it, including patient preferences for that treatment and (b) who performs the procedure. As we discuss below, these sources of heterogeneity make it far more difficult to establish standards of cost-effectiveness and technology assessment.

2.1 The effectiveness of a procedure depends on who gets it

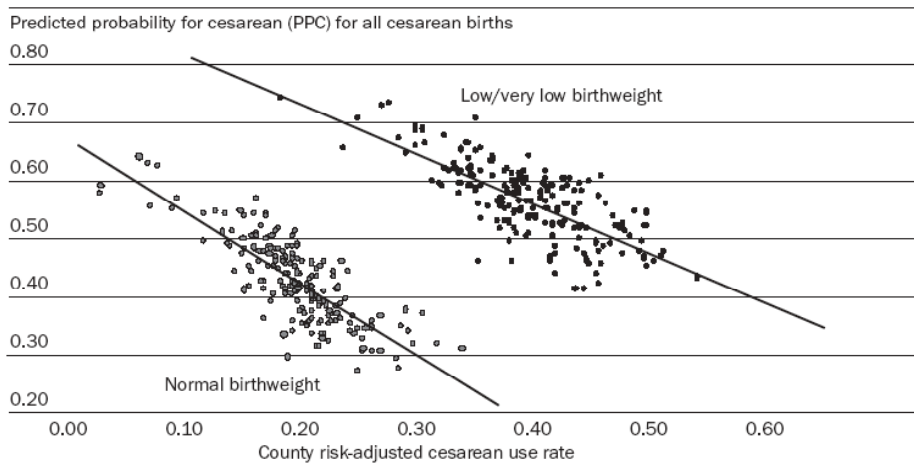
Whether or not a procedure is therapeutic fundamentally depends on a patient’s clinical status. Even for patients within a clinical trial, there are substantial differences in the relative benefits across risk strata, with some patients receiving no benefit (or even harm), while a smaller group achieves a much larger benefit than the average reported for the trial (Hayward, 2005). As Hayward discusses, such limitations could be addressed within CEA approaches and markedly improve allocation. If there are differences across patients in the benefit from a treatment, the answer from a randomized clinical trial (RCT) cannot be used to generalize to the entire population of patients who are potential candidates for a procedure. This dilemma routinely surfaces as a medical technology “diffuses,” and, as part of the diffusion process, is used in patients with declining medical benefit. Unsurprisingly, it is in precisely these murky indications that

procedure use has grown most aggressively. This point is most germane to cost-effectiveness calculations for imaging and diagnostic technologies, where, in principle, it is possible to screen the entire population for every medical condition.

To illustrate, consider the use of cesarean delivery, a surgery with potentially life-

saving benefits in births where the fetus malpresents or has an abnormal heart rate. In such deliveries the procedure meets virtually every cost-effectiveness criteria. However a recent study demonstrates that

Relationship Between Predicted Probability Of Cesarean Birth And Cesarean Rates, Normal Birthweight And Low/Very Low Birthweight, 1995-1998



SOURCE: Authors’ calculations from U.S. natality data, 1995-1998.
NOTES: Relationship between each area’s cesarean rate and the average probability of cesarean delivery among cesarean births in the area (a measure of appropriateness). A ten-percentage-point increase in the cesarean rate for normal-birthweight births (2,500 grams or higher) leads to a fourteen-percentage-point reduction in appropriateness for cesarean births ($p < .001$) and a nine-percentage-point decline in appropriateness for low-birthweight and very-low-birthweight births (less than 2,500 grams) ($p < .001$). Observations are 198 U.S. counties (1995-1998).

Figure 1

increased rates of use are associated with “off-label” motivations such as scheduling for convenience or malpractice fears (Baicker, Buckles, and Chandra, 2006). These motivations cause the procedure to be performed in births where the child is medically less appropriate for it, as shown in Figure 1. The X-axis arrays the largest counties in the United States on the basis on their cesarean rate, and the Y-axis reports a measure of appropriateness for the procedure among births that received cesarean delivery. The negatively sloped lines demonstrate that counties with higher cesarean rates are performing the birth in less appropriate populations; these babies tend to have higher birthweight, longer gestation, and lower incidence of congenital anomalies. While some cesareans are clearly indicated (and trials typically enroll these patients and find positive effects), the more appropriate measure of cost-effectiveness is the value of these marginal procedures done in high-utilization counties in response to (e.g.) high malpractice pressure.

The notion of performing a procedure in patients of declining benefit describes not only the cross-sectional relationship, but also the use of a procedure over time. A particularly good example of this comes from studies involving the use of carotid endarterectomy to treat a narrowing of the carotid artery. The key question in this case was whether the risk of operative mortality was greater or less than the risk of a stroke that might occur in the absence of surgery, and so operative mortality is a critical factor in judging the effectiveness of surgery. In the two major trials from the 1990s, people over age 80 were excluded (see Wennberg, et. al., 1998).

Yet as Figure 2 demonstrates, the use of carotid endarterectomy among the oldest-old population is growing more rapidly than among the young-old (age 65-80), and rates in some regions for the over age-80 group are higher than rates in many communities among the population aged 65-80 (e.g., Las Vegas, where the rate is 3.7 per thousand population over age 80). Despite the high prevalence of carotid endarterectomy among the oldest-old population, there is remarkably little evidence on its efficacy, and is limited to largely observational studies (e.g., Barnett, 2005). In the context of heart-attack treatments, Chandra and Staiger (1997) suggest that intensive management of heart attacks in 1994/95 (as proxied by the receipt of cardiac catheterization) increased one year survival by 17 percentage points for patients aged 65 to 80, but only by 1.5 percentage points in those aged over 80.

Technology assessment explicitly recognizes the above notion of declining medical benefit, and can in principle inform the use of cesarean delivery in relatively healthy births, or the use of endarterectomy in the over-80 population. The difficult part is gathering evidence, since for many in these populations the norm is to provide such treatments, leading to resistance among both physicians and potential patients in creating randomized trials where control groups

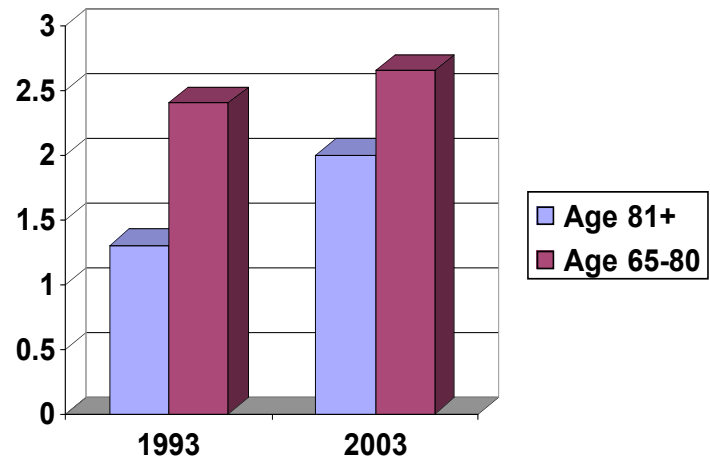


Figure 2: Carotid Endarterectomy Rates Per 1,000 Medicare Enrollees, by Age, 1993 & 2003
Source: Dartmouth Atlas, Medicare Claims Data

would not receive the intervention. Observational trials may provide some guidance, particularly those using “quasi-randomization” methods such as instrumental variables, but even these are not a complete substitute for a fully randomized trial.

Another difficulty comes in the evaluation of “preference sensitive” treatments (in the sense of Wennberg, Fisher, and Skinner, 2002) where some patients may rationally choose the treatment while others do not. For example, some women prefer to be treated with breast-sparing surgery followed by radiation therapy for breast cancer, while others prefer mastectomy. Randomized trials may reflect average satisfaction, costs, and clinical outcomes of the “treatment” and “control” groups, but do not reflect the more relevant measure – which is the cost-effectiveness of (say) mastectomy among those patients who would prefer the mastectomy option. More generally, providing patients with options that are closely aligned to their intrinsic preferences yields value not generally measured in conventional cost-effectiveness studies.⁴

A more problematic case of heterogeneity in patient preferences comes from a situation where most patients lose from the treatment. Topol (2006) presents evidence that the drug nesiritide approved for decompensated congestive heart failure is on average, deleterious for health. Despite this fact, patients with particularly severe CHF may rationally chose treatment, either for symptom relief or for a small chance of improving, given that the alternative is continued severe illness or death. This desire of terminally ill patients to try nearly anything, regardless of cost-effectiveness criteria, along with liberal insurance company policies strengthened by relatively low coinsurance rates in the U.S., may explain the existence of cancer treatments with very high prices but little proven clinical value (Bereneson, 2006).

2.2 The effectiveness and costs of a procedure depend on where the procedure is performed and who does it

The effectiveness of a given technology (and thus its effectiveness compared to alternative treatments) depends critically on the skills of health care providers, whether because of volume effects or other factors more difficult to identify. In the trials of carotid endarterectomy discussed above, patients were randomized into the trial (conditional on being under the age of 80), but there was a stringent process for choosing which hospitals could participate in the trials:

The ACAS [Asymptomatic Carotid Atherosclerosis Study] trial used a 2-step selection process: hospitals had to submit mortality and morbidity data; if the center had acceptable results, individual surgeons within the institutions submitted data showing a combined perioperative event rate of less than 3%. Additionally, ACAS had ongoing evaluations of perioperative events; if more than 1 perioperative event occurred, an institutional audit was undertaken to evaluate eligibility for further participation in the trial. (Wennberg et al, 1998; p. 1278)

Not surprisingly, the Wennberg et. al. study found quite dramatic differences in perioperative mortality depending on whether the procedure was performed in one of the original trial hospitals (1.4 percent), in a non-trial hospitals with high volumes of endarterectomies (1.7

⁴ A patient’s anxiety and risk aversion are closely related but reference distinct psychological phenomena. Anxiety requires the presence of uncertainty, but it also requires a sufficiently long time-interval before the uncertainty associated with a treatment or disease is resolved. Patients will discount the resolution of the uncertainty at different rates based on their time-preferences. These three sources of variation in preferences will result in two clinically identical patients perceiving very different benefit from a treatment.

percent), or hospitals with low volumes (2.5 percent). In other words, procedures that are highly cost-effective in academic medical centers may not be so in community hospitals. Another study by Chandra and Staiger (2007) found that regions that specialized in treating heart attack patients with intensive management obtained better results with this therapy than in regions relying on medical treatments, with cost-effectiveness ratios ranging as much as four-fold, at least with regard to Part A (inpatient) hospital expenditures.⁵

2.3 Technology assessment focuses on one treatment, but may fail to consider other interventions

Cost-effectiveness analysis reveals the effectiveness of a given intervention holding the effect of all others constant. In practice however, this may not be true: the use of a given procedure may crowd out the use of another equally efficacious one. If the diffusion of angioplasty slows the diffusion of beta-blockers, we would have overstated the return to angioplasty. Similarly, one of the reasons for why some regions of the US have a higher benefit to performing PCI is that they perform poorly in the medical management of patients (at the level of the hospital referral region, the correlation between catheterization rates and beta-blocker use was -0.31 in 1994/95). For this reason, Stukel et. al. (2005) found that regions with aggressive medical management of heart attacks (e.g., beta blocker and aspirin use) experienced similar long-term outcomes to regions with aggressive surgical management. Again, this finding points to outcomes assessment – comparing the overall outcomes of regions or hospitals that may adopt quite different strategies to treating patients with similar clinical characteristics.

Specific technology rules or even “process-based” quality standards such as HEDIS measures rarely capture differences in actual quality of care across providers, perhaps because important factors that result in improved outcomes (such as better coordination, effective counseling to achieve adherence with exercise and medication management) are not captured but are more important in achieving good outcomes. One recent study showed only modest correlations between the hospital Medicare quality measures and actual risk-adjusted outcomes, suggesting natural limits on how much measured process can capture true quality of care (Werner, 2006).

Finally, technology assessment and cost-effectiveness analysis may not capture dimensions of health care practice that appear to be most important in driving both cross-sectional differences in spending and growth in spending over time. The nearly two-fold differences in longitudinal costs observed across academic medical centers in the care for patients with AMI are largely due to discretionary decision-making about how frequently patients should be seen, how often similar patients are referred to subspecialists, whether patients are cared for in the hospital, and the intensity of diagnostic testing and imaging procedures (Fisher, et. al., 2004). Differences in spending aren’t due to “what” is provided (PCI vs. aspirin), but “how” (the labor and associated services that are bundled with it in higher cost systems). We consider this point in more detail in the next section.

3. Hospital-specific expenditures and outcomes in the treatment of heart attacks

We illustrate the empirical importance of variations in expenditure and outcomes by focusing on the treatment of AMI in the Medicare population from 1992-2004 for several of the

⁵ We have demonstrated the presence of heterogeneity in treatments effects that vary across providers and patients. One solution is to perform cost-effectiveness analysis in two trials, one randomizing over patients and another over providers, to document variation in provider skill and in the match between patient and physician.

larger hospitals where sample sizes allow accurate assessment of both costs and outcomes. This case is closest to traditional technology assessment, since an AMI is relatively well defined and there are well-established and validated methods for risk adjustment, even in administrative data (Krumholz, et. al., 2007), and general agreement on the validity of the outcome variable, mortality.

3.1 Data and Measurement

We began with a 100 percent sample of Medicare Part A claims data from 1992-2004 were merged with the Medicare Denominator File through 2005 to create a longitudinal cohort of fee-for-service enrollees age 65 or over coded with a new acute myocardial infarction. Patients with a code of “old MI,” or those identified from the panel data as having had an AMI previously, were excluded from the sample.⁶ Overall there were 3,012,934 valid AMI events. In this study, we focus on a limited subset of the larger sample, consisting of larger hospitals where there are at least 250 heart attack patients in any given year. This was done solely to ensure sufficient sample size for statistical precision. Thus we consider outcomes and expenditures for these 25 largest cardiac hospitals, comprising 119,587 AMI patients. In our analysis, we further consider a subset of five hospitals that provide a range of contrasting expenditure and outcomes measures.

We considered just Medicare Part A expenditures (excluding patient coinsurance and Medicare Part B expenditure; omissions that cause us to overstate the effectiveness of spending), correcting for inflation using the US implicit price deflator with all results expressed in 2004 dollars. To adjust for both secular and cross-sectional differences in health status, we adjust survival rates and expenditures for a variety of comorbidities (diabetes, diabetes with complications, pulmonary disease, liver disease, liver disease with complications, dementia, non-metastatic cancer, metastatic cancer). Also included were age-sex-race effects consisting of 5 age categories (65-69, 70-74, 75-79, 80-84, and 85+) interacted with sex and with two race variables (black and nonblack), and the type of MI (inferior, anterior, subendocardial, and other). Controlling for the type of infarction is important, since the fraction of the less serious subendocardial (or non-q wave) heart attacks rose during this period because of more sensitive enzyme tests for the presence of an AMI. All estimated survival and expenditure measures are expressed in terms of the representative patient with average characteristics for the hospital sample during the entire period of analysis. (These may differ from the general population of AMI patients, particularly those admitted to smaller hospitals and thus excluded from the sample.) Hospitals will differ both with regard to their initial adjusted survival (and expenditures), and with respect to changes over time in these variables, but our approach will, as far as possible, ensure that the results reflect hospital-level practice patterns rather than differences in patient characteristics.

There are a variety of approaches to treating patients with AMI, and we have limited information for each hospital on their technology adoption. During 1994/95, the Cooperative Cardiovascular Project (CCP) conducted a chart review of roughly 160,000 AMI patients regarding the use of treatments such as β blockers and aspirin within 24 hours of the AMI.

⁶ Other exclusions included if patients were enrolled in an HMO at the time of the heart attack. This data description corresponds closely to the discussion in Skinner et. al., 2006. Because of our focus on hospitals, we use only data from 1992 onward because data from 1991 and earlier provides just a 20% sample, which is too small for precise estimates among hospitals.

Aspirin reduces platelet aggregation and has been well-established in reducing the risk of mortality following AMI. Beta blockers are an inexpensive drug that by blocking the beta-adrenergic receptors reduces the demands upon the heart, and have been known since the mid-1980s to be effective in reducing post-AMI mortality by 25 percent or more (Yusuf, et. al., 1985). But compliance in the use of β Blockers has lagged among many states, even as late as 2000/2001 (Jencks, et. al., 2003).

Hospital-level utilization rates were estimated for each hospital in the sample and merged with the Medicare claims data by provider number. As well, we also consider the fraction of AMI patients in the hospital being treated with PCI (percutaneous coronary interventions) which includes both angioplasty and stents. (Similar results were found when a more inclusive measure of PCI plus CABG (bypass) rates was included.) PCI has a somewhat mixed record with regard to technology assessment. It is well established that PCI within the first 12 or 24 hours of the AMI leads to better survival rates (Keely, Boura, and Grimes, 2003), but there is no evidence that a PCI done subsequently, for example to clear an occluded artery, has any impact on survival (Hochman, et. al., 2006) although it may improve functioning. Nonetheless, it does appear that the rate of primary angioplasty is highly correlated with the rate of overall angioplasty, and so we use the 30-day PCI rate as a rough proxy for expensive but potentially useful treatment.

3.2 Results

The first column in Table 3 provides summary measures of both cost changes and outcomes changes for the sample of the 25 large hospitals. (Results are similar to those for the entire sample.) Since 1992, there has been a decline in the one-year mortality rate equal to 4.9 per 100 AMI patients. As noted earlier (Skinner, et. al., 2006), most of this decline occurred in the early to mid-1990s, more recently mortality gains have largely stagnated. Risk-adjusted inpatient Medicare expenditures rose by \$7,397 during this period. Assigning a conversion factor based on the Cutler and McClellan (2001) analysis suggests that, over this time period, there was a remarkably favorable cost-effective ratio of \$12,455 per life year. It may appear that we are calculating conventional cost-effectiveness ratios, but recall that these measures reflect the multi-

	Average*	Hosp A	Hosp B	Hosp C	Hosp D	Hosp E
Adj. 1 Year Mortality, 1992	0.346	0.366	0.415	0.326	0.361	0.291
Adj. 1 Year Mortality, 2004	0.297	0.250	0.305	0.289	0.356	0.294
Mortality Diff.	-0.049	-0.116	-0.110	-0.037	-0.005	0.003
Adj. 1 Year Expenditures, 1992	19,991	14,785	16,492	22,961	18,799	15,425
Adj. 1 Year Expenditures, 2004	27,388	21,904	23,494	41,002	28,717	23,326
Expenditure Diff.	7,397	7,119	7,001	18,041	9,918	7,901
PCI Rate, 1992	0.27	0.33	0.17	0.23	0.23	0.43
PCI Rate, 2004	0.47	0.59	0.43	0.42	0.35	0.53
Beta Blocker, 1994/95	0.67	0.64	0.65	0.76	0.55	0.35
Aspirin (%), 1994/95	0.88	0.82	0.91	0.95	0.85	0.85
Effectiveness ratio	\$12,455	\$5,064	\$5,251	\$40,231	\$163,633	**

* Averaged over all 25 hospitals with at least 250 AMI patients in each year 1992-2004.

** Not defined

Table 2: Hospital-Specific Measures of Mortality Outcomes and Medicare Expenditures for Five Large Hospitals, and Averages Across Twenty-Five Hospitals, 1992-2004

dimensional process of care, and the determinants of costs, and these are not likely driven primarily by the introduction of one or two technological innovations. Thus they really must be interpreted as associations rather than the “causal” effects as in traditional cost-effectiveness analysis. On average, there was also a sharp rise in the use of PCI within 30 days of the AMI

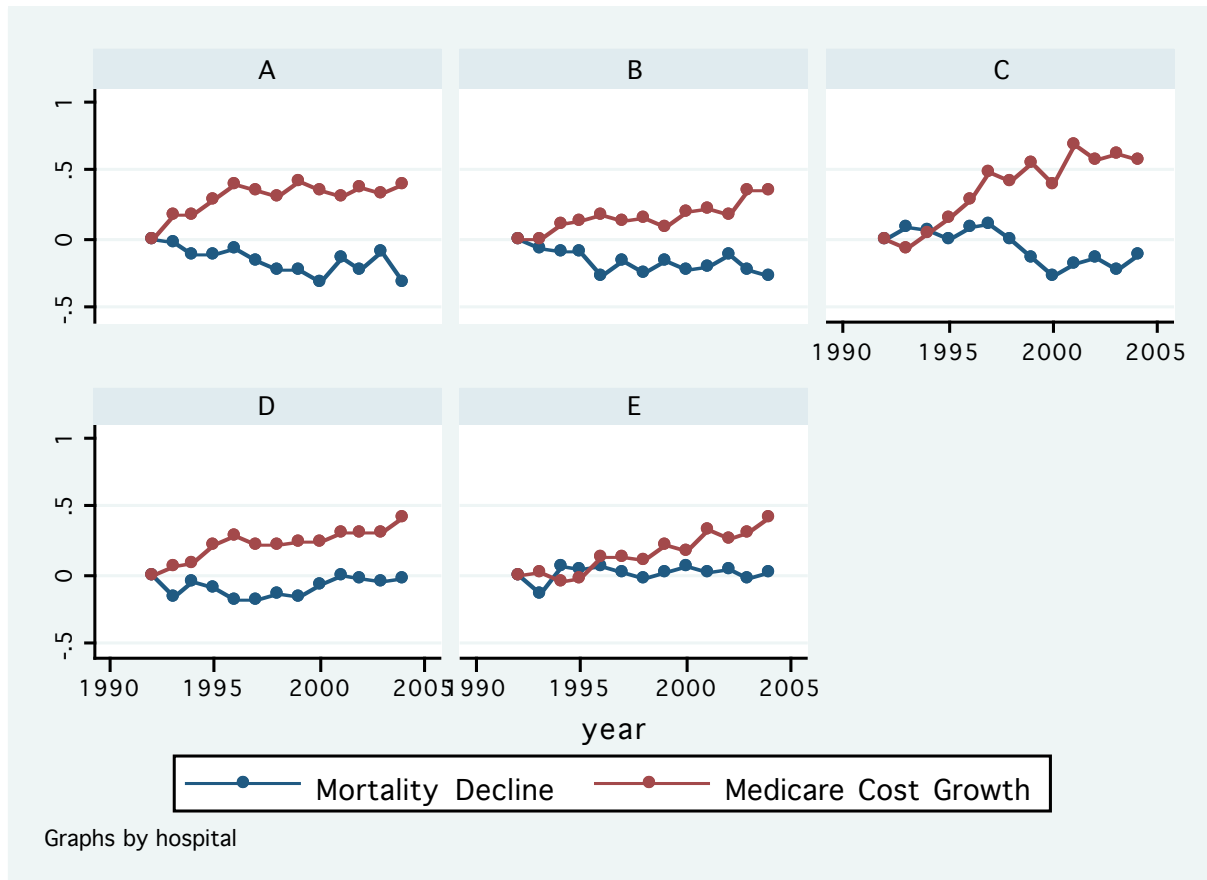


Figure 3: Proportional Mortality Decline and Log Growth in Medicare Expenditures (1992 = Baseline) For Five Hospitals, 1992-2004

Notes: Each hospital has at least 250 AMI admissions per year for each year. Hospitals are ranked from A, among the most cost-efficient (a cost-effectiveness ratio equal to \$5,064) to E, the least cost-effective, with an undefined cost-effectiveness ratio.

(from 27 to 47 percent of patients). In 1994/95, β blocker use among ideal patients was 67 percent, and aspirin at discharge was 88 percent. These are somewhat higher than the overall average in the US for the same time period.

The 5 hospitals were ranked by their own cost-effectiveness ratios, again defined as the change in expenditures divided by the change in risk-adjusted life expectancy. For the 5 hospitals chosen, their individual effectiveness ratio ranged from one that was highly favorable (A), just \$5,064 per life year (Table 2), to a ratio of \$163,633 for Hospital D, and to an undefined ratio for the least effective hospital (E), because expenditures rose while mortality did not change. Figure 3 shows both proportional mortality changes (that is, the fractional decline in mortality relative to the baseline mortality in 1992) and the log change in expenditures, again for

each of the five hospitals. There are dramatic differences in the progress of each hospital during this period, with Hospital A demonstrating a far more favorable growth pattern than Hospital E.

While PCI rates grew in all five hospitals, there does not appear to be a strong correlation between either levels or rates of PCI growth, whether among the five hospitals, or more generally among all hospitals. As it happens, Hospital E was quite low with regard to β blocker use in 1994/95, but paradoxically the hospital's initial mortality rate in 1992 was quite low as well. This underscores also the importance of rewarding performance not just on the basis of changes in outcomes and costs, but also with regard to the absolute levels of quality. It is possible that Hospital E effected productivity changes prior to 1992, thus explaining why its initial mortality rates in 1992 were so low. Presumably the optimal approach to compensating hospitals for quality involves rewarding both on the basis of the level of quality and expenditures, as well as changes in quality and expenditures.

As can be detected by an examination of Table 2, and further analysis of the data, the patterns of change in mortality and expenditures are not well explained by conventional measures of technological efficiency, whether beta blocker and aspirin use in 1994/95, or by contemporaneous PCI rates (or bypass surgery rates, not shown). The correlation between β blockers and mortality change is significant ($\rho = -0.22$) but in general, the other correlations are modest, whether the rise in PCI and mortality (-0.11 , not significant) or the near zero correlation between the increase in PCI and the increase in expenditures. The mystery is therefore why some hospitals seem to have adopted the low-cost technology while others simply appear to have cost more money without above-average gains in outcomes. Still, one can imagine reimbursements systems that reward hospitals with lower-than-average cost (or cost growth) or better-than-average survival (or survival improvements). More generally, some of the cost saving could be used to implement quality improvement efforts in hospitals that consistently fall behind with above-average spending and below-average survival. As well, we would expect that these "high-powered" financial incentives would create a strong environment to improve the quality of risk adjusters in evaluating health outcomes.

4. Practical Considerations in Outcomes Assessment

While this empirical exercise is illustrative, it still falls short for at least several reasons. First, it is clearly not sufficient to measure AMI outcomes alone, given that heart attack patients account for only a small fraction of total health care costs.⁷ Second, any serious efforts to address the current inefficiency in health care will require changes in the structure of incentives, a difficult political and technical challenge given uncertainty about future health care reform. This brings us to our final point which we believe deserves particular emphasis – that any effort to close gaps in efficiency will require defining the individuals or organizations that can serve as a locus of accountability for measuring and improving outcomes. Each of these issues is addressed in turn.

⁷ In the Medicare fee-for-service population, there are at most 250,000 AMI patients in a given year. Assuming that the incremental cost of such patients is about \$30,000 per patient (a high estimate), the aggregate amount of expenditures for AMI patients amounts to roughly \$7.5 billion, or less than 2.5 percent of total Medicare expenditures in 2005.

4.1. Challenges in measuring outcomes

It is relatively straightforward to articulate a plausible set of measures for acute episodes such as acute myocardial infarction, elective joint replacement or pneumonia. These include mortality, functional status, and total episode costs using standardized prices, for example. But it is more difficult to envision a comprehensive set of outcome measures to capture the overall health of the target group of patients. (We discuss below how to define these specific groups.) For the general population, measuring outcomes and costs accurately would require substantial data collection, including, presumably electronic medical record system, and even the administration of surveys such as the SF-36 or more recent improvements on measuring health, through the PROMIS database⁸. At the same time, it might be reasonable to select samples of the population with different tracer conditions and specific relevant measures, for example the general population (screening and satisfaction with care); diabetes (health status, functioning, predicted risk of death); cancer (mortality and quality of life).

The use of outcomes measures should not crowd-out the use of process measures (e.g., HEDIS measures), since individual organizations have a strong incentive to monitor and reward these dimensions of health care quality given their contribution to improved outcomes. Outcomes assessment would create incentives for systems to implement process measurement to better inform their understanding of the factors that improve health status. For example, outcomes measurement provided the impetus for the collaborative efforts to understand how cardiac bypass (CABG) mortality varied across hospitals and to introduce changes in process to improve those outcomes. (Hannan, et. al., 1994; O'Connor, et. al., 1996). A further question is whether future payment systems should reward quality based only on outcomes measures, or on some combination of (imperfectly measured) outcomes and (more accurately measured) process measures?

What about measuring outcomes in the case of “preference sensitive” care, for example the sample of patients receiving treatments for prostate cancer, BPH, or breast cancer where a variety of treatment options exist? In this case, process measures – whether the patient received appropriate assistance in arriving at a well-informed decision – would be an integral component of evaluating performance.

Although the measurement challenges are substantial, the difficulties should at least be considered in the context of the potential importance of creating incentives to improve the overall efficiency of care. The current initiative by the National Quality Forum to develop a measurement framework for efficiency (encompassing both outcomes and costs) underscores the value that a broad group of stakeholders encompassing purchasers, health plans and providers places on a measurement system that could reliably capture costs and outcomes to inform judgments about efficiency over prolonged episodes in defined populations. While we recognize the myriad possibilities of “gaming” such a system, we still believe that accurate health outcomes measurement represents a very important future goal for any rational health care system. One cannot pay for performance without measuring performance accurately.

⁸ This is a network of researchers who have developed inclusive databases on questions designed to measure patient outcomes. See <http://www.nihpromis.org/>.

4.2. Different health care systems, different incentives

The premise of this paper has been that under any proposed reform, severe challenges will exist to ensure high quality care while at the same time restraining growth rates in health care costs. Here we consider the incentives systems under fundamental health care reform, by which we mean universal coverage (or near universal coverage) in the U.S. Consider three broad categories: an incremental expansion of coverage to the uninsured, for example as proposed in Massachusetts and California; single-payer national health insurance; and insurance vouchers.

4.2.1 Incremental expansion

This approach preserves to the greatest extent the status quo, and provides universal coverage by supplemental government programs with mandatory “buy in” provisions that force enrollees to pay a premium based on their income, perhaps accompanied by a tax on firms that don’t provide health insurance. The government would need to provide additional funds to subsidize the additional coverage not covered under the quite modest premiums. It’s not entirely clear that all states could afford to adopt such a program, particularly those such as Texas with a large fraction of its population uninsured, but it is clear that this incremental approach leaves unaltered the existing U.S. trifurcation of health care financing (government, private insurance and out-of-pocket), with an expanded role for government subsidies. Thus it seems unlikely that there would be any additional pressure to adopt new approaches to outcomes assessment under this type of reform.

What incentives exist currently to adopt outcomes assessment? There is increasing interest in “pay-for-performance” incentives structures, but these initiatives often focus on adding specific technical process measures and making minor additional payments for improved care. But the problem of efficiency (reducing costs while improving quality) is largely ignored in current efforts. In cases where there is real cost saving, whether through prevention or low-cost options to expensive and highly remunerated surgery, providers actually lose money, as in the case of a back pain clinic steering patients to low-cost rehabilitation before sending them in to the hospital for more expensive diagnostic tests and potential back surgery (Fuhrmans, 2007). A notable exception is Medicare’s Physician Group Practice demonstration (CMS, 2005) in which CMS offers physician groups a share of any savings below a projected target growth rate if they also meet quality targets. Still, there is little incentive to adopt technology assessment, much less outcomes assessment. It is harder to imagine the current system persisting given the likely sustained growth in health care costs, and if anything expanding coverage to the uninsured in this incremental way could only hasten a future fiscal crisis.

4.2.2 Single-payer national health insurance

This approach represents the most dramatic transformation of health care in the U.S. One good example of a proposal comes from the Physician’s Working Group for Single-Payer National Health Insurance (PWG, 2003) which has proposed a hybrid system in which care is provided either through a discounted fee-for-service plan, like that in Canada, or through strengthened managed care organizations funded through capitation or global budgets. Hospitals would be highly regulated, for example they would not be allowed to engage in capital investments without prior approval from the central insurance board, and for-profit hospitals would be bought out and converted to not-for-profits. The dynamics of such a program would be not dissimilar to the Canadian or British systems, where the primary limitations on spending

growth would arise from slowly growing budgets limited by the appetite of American taxpayers for tax hikes.⁹

The current PWG proposal does not focus on lagging quality and how that might be addressed under a single-payer system. The coordination of health care spending under a single umbrella could in theory allow for the implementation of outcomes assessment and rewards, either in the form of per capita reimbursements or “good performance” bonuses, for high-value low-cost care.¹⁰ As well, the existence of a single-payer system would presumably enhance the ability to develop universal health care cost and quality measures that can span different providers – that is, assuming that a universal information technology system could be developed. What is unknown about a hypothetical single-payer system is what form incentives for improved quality or reduced costs would take. The Physicians Working Group seeks to remove health care from the tainted hand of the market, raising questions about how one should reward clinics or groups who manage to provide high quality care at lower costs. England has experimented with rewarding physicians for improving the process quality of care, but extending such a program with regard to outcomes – for example, paying more on the basis of lower risk-adjusted mortality – might be politically more difficult. A further challenge is creating provider organizations large enough to measure outcomes accurately, but small enough to allow for organizational agility, a topic we address below.

4.2.3 A voucher system for health insurance

Another option to provide universal coverage is the use of a health insurance voucher that is sufficient to pay for at least a minimal coverage level. Emanuel and Fuchs (2005) have proposed one such plan, which they refer to as a Universal healthcare voucher (UHV). In this approach, every citizen receives a health insurance voucher which they can then use to purchase insurance coverage. Thus insurance companies play a central role in competing for vouchers by providing desirable insurance plans. (By contrast, under single-payer proposals, health insurance corporations could be slated for extinction.) One might expect such a structure to provide a particularly fertile environment for outcomes assessment. In marketing their individual plans to consumers, insurance companies will be particularly sensitive to the overall costs of their enrollees, but also should seek reliable measures of outcomes to help attract additional enrollees. After all, consumers presumably care about the functioning and mortality outcomes of plan enrollees, as well as the cost. The real challenge is to encourage marketing efforts that focus more on these objective measures of outcomes, and less on the availability of consumer amenities and valet parking.

Missing among these options are reforms based on consumer-driven health care such as the expansion of health savings accounts and tax incentives to purchase low-cost insurance plans. These plans are typically coupled with catastrophic health care insurance that picks up costs that exceed a certain amount. One could certainly imagine such reforms as integral parts of

⁹ Given that the U.S. spends such a large fraction of its GDP on health care, and its taxpayers have historically expressed a great aversion to heavy tax burdens, resistance to additional tax hikes would presumably be effective in restricting future expenditures growth.

¹⁰ The PWG proposal explicitly rules out incentives based on saving money, but presumably saving money while improving health care quality would not be frowned upon.

incremental reform (4.2.1), for example, the implementation of health savings accounts currently proposed under the Bush Administration. Or they may be a central component of competition for low-cost health insurance plans under a voucher reform (4.2.3). But by itself, consumer-based health care does not provide the universal coverage under fundamental health care reform.

Also missing in this discussion is the importance of creating organizations to translate incentives from theory into practical action. Measurement of outcomes and costs requires identifying both the responsible provider or providers and the patients (or population) whose care is to be measured. While current efforts in the U.S. marketplace tend to focus on individual physicians and institutional providers (e.g. hospitals, nursing homes), these have serious limitations when trying to examine outcomes: sample sizes for individual physician practices are small,¹¹ while any serious illness, whether acute or chronic requires the care of multiple physicians and often multiple institutional settings. Indeed, the most serious gaps in quality are a consequence of flawed transitions and poor coordination for such patients and differences in costs across hospitals and regions largely reflect how many institutional and professional resources are brought to bear on the care of similar patients (Fisher, 2004). These would continue to exist under any system of care, for example single-payer health insurance and vouchers. For this reason, we consider a simple “default” approach to creating organizations that may easily respond to incentives present under outcomes-based incentive schemes.

4.3. Translating incentives into practice: The Accountable Care Organization

Improving efficiency will require identifying entities with adequate sample sizes that can take responsibility for integrating care over time and across different providers. We label such entities Accountable Care Organizations (ACOs). Large multi-specialty physician group practices (such as the Mayo Clinic) or traditional HMOs (such as Kaiser), could clearly serve as a locus of accountability for longitudinal costs and outcomes, but these represent only a small share of the current market - and most physicians remain in solo or very small group practices. Some have suggested that individual physicians -- in the role of a medical home -- could help improve coordination, but these models have yet to be tested in practice and how they would foster integration across hospital, nursing home or other post-acute settings is far from clear.

An alternative approach that provides for the measurement of longitudinal costs and outcomes was proposed by Fisher et al (2006), based in turn upon the medical staff proposal put forward by Welch and Miller (1989, 1994). The proposal was based upon several empirical observations: (1) almost all physicians work within or around a single hospital and can be directly affiliated with that hospital using Medicare claims data (Bynum et. al., 2006), thus representing a “virtual” multi-specialty group practice; (2) the patients cared for by these empirically defined medical groups can be identified through claims; (3) over a one-year period, most of the care for these patients is provided by the empirically defined medical group or a referral hospital and its staff that are readily identified. These empirically defined hospital / medical staff groups thus provide care to relatively large and stable populations (providing statistical precision in both outcome and cost measures) and the hospital / medical staff group is

¹¹ In analyses that assigned Medicare beneficiaries to their predominant care physician, only half of physicians were assigned any beneficiaries (pathologists provide no direct patient care) and only 16% had one 25 or more patients assigned. (Fisher 2006).

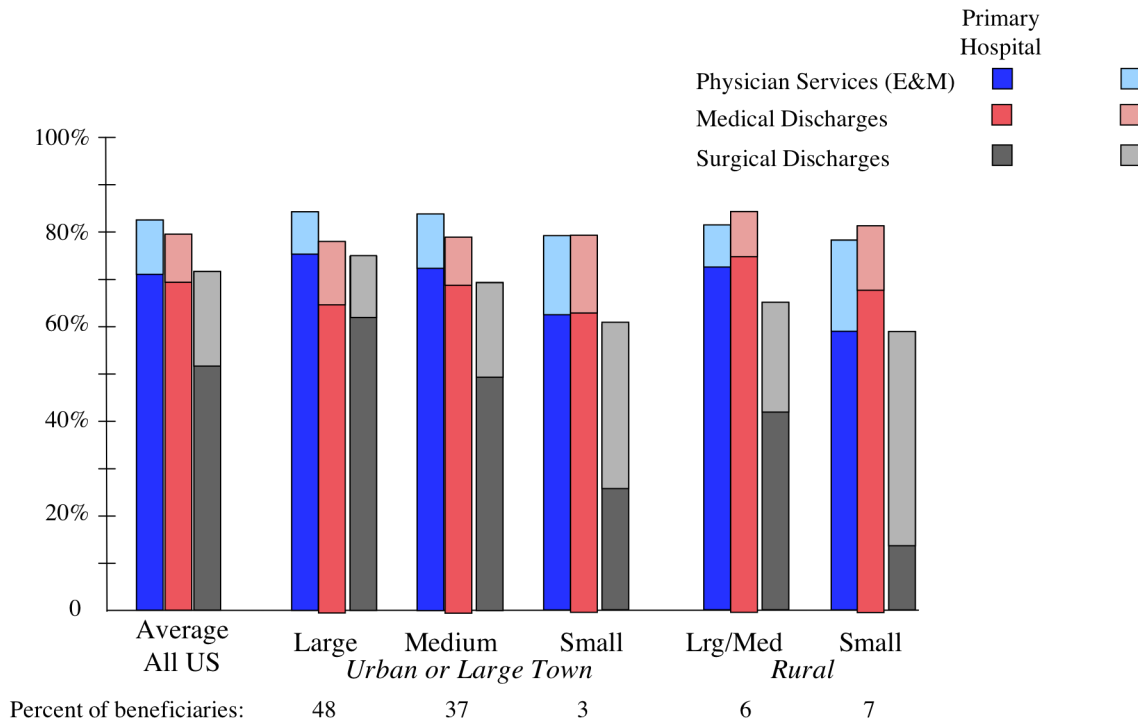


Figure 4: Percent of beneficiaries' care at assigned primary or secondary hospital

responsible for most of their care, thus providing a potential basis for defining an entity that could conceivably serve as an ACO.

Fisher and Gottlieb (2006) described an empirically defined ACO by (Step 1) assigning physicians to hospitals and (Step 2) assigning patients to hospitals (the two steps are independent). The resulting “network” of physicians and patients assigned to a particular hospital may be treated as an ACO and provides a coherent quantity for rewards and measurement, though we reiterate that no formal contracting agreement binds the physicians or patients assigned to an ACO. For the physician part of the assignment, they assign MDs to hospitals using two assignment rules: MDs with inpatient work were assigned to the hospital where they provided care to the greatest number of Medicare beneficiaries. MDs with no inpatient work were assigned to the hospital where the plurality (usually majority) of patients they billed for were admitted. Using this simple algorithm they were able to assign 95% of MDs to a single acute care hospital in the US.¹² On the patient side, they used two sequential assignment rules: assign each patient to their predominant care physician (primary care MD or medical subspecialist), and assign patients to primary hospital based on their *physicians'* assignment. (An additional assignment was made to a secondary referral hospital.) This algorithm allows patient X to be matched to physician Y, and through that link, to hospital Z (even though the patient may or may not have been treated as hospital Z). This algorithm allowed the assignment of 93% of ambulatory care patients to an acute care hospital.

Figure 4 illustrates the coherency of the ACO definitions. The figure distinguishes

¹² They found that 62% of physicians work at only one hospital, and an additional 28% (comprising multi-hospital physicians) perform their work at an assigned hospital.

between primary and secondary hospitals. Focusing on the averages for all US hospitals, 70 percent of the evaluation and management services provided to beneficiaries assigned to a hospital is provided by the physicians within their primary hospital (as created by Fisher and Gottlieb), and about another 10 percent is provided at a secondary referral hospital. So more than 80 percent, on average, of the evaluation and management services, are provided at the hospital to which they've been assigned.¹³ Looking across columns, 90 percent of the Medicare beneficiaries are in systems that have a very high degree of coherence in large urban, large medium, and large rural hospitals. Clearly, even a simple rule for defining an ACO produces a relevant unit for analysis. These "default" organizations could of course be altered by the providers themselves, for example if efficient providers that meet certain minimum size requirements break away to form their own ACO-hospital alliance, but the value of having default assignments cannot be overestimated.

While it is feasible to construct physician groups empirically and the paths to both measurement and payment reform can be at least be imagined, the political and cultural barriers could turn out to be substantial. Most physicians remain in small group practice and have long traditions of autonomy and individual responsibility that may lead to resistance to collaboration and shared accountability. Although counterexamples can be identified (Cortese, 2007), physician-hospital relationships have been increasingly strained in some health care markets in recent years (Berenson, 2006). Encouraging patients to remain aligned with a single care system may bring back memories of early attempts to promote capitation. And any substantial movement toward shared savings approaches would require overcoming legal barriers to gainsharing (Wilenski, 2007).

5. Conclusion

Technology assessment is a necessary component of any health care reform, but it is probably not a sufficient one. As we have argued above, health care systems, whether hospitals or provider groups, may exhibit vastly different levels of risk-adjusted outcomes and expenditures, even when they are using apparently similar technologies. These differences suggest to us that outcomes assessment of specific health care systems is a worthwhile goal for any future health care system. It may be the case that some health care reforms providing universal coverage for U.S. citizens are better suited for outcomes assessment than others, but we would further argue that any reform *should* encourage concerted use of outcomes assessment both to improve quality and to control health care costs. There is no reason why a system even as flawed as our current fee-for-service program could not provide a supporting environment for the development of ACOs that can respond financial rewards to providing efficient health care.

This paper has emphasized the difference between technology assessment and outcomes assessment. In particular, many of the problems inherent in technology assessment, such as the use of procedures on patients not originally included in -- and by physicians not originally eligible for -- randomized trials, are not critical defects in outcomes assessment, where an excessive use of "off-label" procedures or a poorly performing surgeon would be revealed through excessive mortality or poor satisfaction with care. Similarly, a greater reliance on outcomes assessment can sidestep the necessity of making explicit choices regarding the use of

¹³ The coherency of the ACO is lower for medical admissions (but still acceptable), while surgery is lower still. Because surgery is often referred to other specialties, there are many categories of surgery for which a patient may be referred to outside even a secondary referral hospital.

procedures with unfavorable cost-effectiveness ratios. ACOs would not be barred from using such procedures, they would simply need to find a revenue source (such as higher premiums) to pay for such procedures. Presumably most ACOs competing for enrollees on both price and quantity would simply eschew such treatments rather than complain about why they can't provide them to willing patients.

That said, we are probably guilty of overstating differences between technology assessment and outcomes assessment, since they both share the same framework and ultimately the same goal to provide high-value care. Certainly, low cost-effectiveness even among ideal patients is a strong signal to the responsible organization (whether a single payer or an ACO) that the procedure is best not provided for their enrollees. And outcomes assessment only strengthens the demand for cost-effectiveness studies that address "off label" patients or different types of physicians. For example, one could imagine a greater emphasis on randomized trials that correspond more closely to the types of practices or hospitals where the treatment is provided. Or randomized trials that focused on less "technical" issues such as whether 3-month follow-up visits provided better health outcomes than 6-month follow-up visits.

Finally, the pursuit of outcomes assessment approaches could distract those in the United States from reaping the obvious gains from technology assessment. That is, before chasing after the "first-best" of outcomes assessment, perhaps policy makers in the U.S. should follow the lead of the U.K. in providing more financial teeth for technology assessment in the Medicare and Medicaid programs, along the lines of the National Institute for Health and Clinical Excellence (NICE) (Pearson, 2007).

One real limitation of this paper is that we have not shown conclusively that outcomes assessment could have a real impact on *growth rates* in health care expenditures. This is largely uncharted territory, but since our Accountable Care Organizations (ACOs) already exist by default, it is possible to identify ACOs with either higher-than-average or lower-than-average growth rates in expenditures. For example, data from Fisher (2006) suggest that growth across ACOs in spending (defined as the change in expenditures divided by average initial expenditures) varied from 10% over 4 years for the slowest growing quintile of medical groups to 46% over 4 years in the highest growing quintile. Under a shared savings model, with an annual growth rate target of (say) 15%, the bottom quintiles could have received a share of the savings to Medicare achieved by remaining below the target, with rewards funded by payment reductions assessed on ACOs growing at higher rates. (This shared savings approach was also used successfully to constrain the growth in per-capita hospital costs in Rochester NY.) Thus even in the short term, a shared savings approach could provide an interim solution as efforts to more fully develop prospective payment approaches.

One puzzle is why traditional technology assessment measures tell us relatively little about the performance of individual organizations. As noted in our analysis of the AMI data, some hospitals appear to provide low-cost high-quality care while other hospitals appear to be struggling, providing high-cost low-quality care. We don't have an explanation for this variation in outcomes, only that we can reasonably rule out that such differences are the consequence of traditional technology adoption – all of the hospitals in our sample had, by 2004 embraced PCI as a common treatment for AMI. But perhaps there are other, less obvious innovations that we are not measuring very well, but which matter for outcomes and costs. The variations in per-beneficiary costs at both the regional and hospital levels are strongly associated with the size and composition of their physician workforces relative to the size of the population they serve, with

the greatest differences found in the per-beneficiary inputs of specialists (e.g., Goodman, 2006). Simply paying organizations for performance may not of itself be enough for these groups to adopt best-practice or reduce their costs. But it will certainly cause them to pay more attention to how to improve outcomes and satisfaction for their patients while reducing growth in expenditures.

References

- Ashenfelter, Orley and Michael Greenstone. "Using Mandated Speed Limits to Measure the Value of a Statistical Life," *Journal of Political Economy*, 112 (2004): S226- S267.
- Baicker, Katherine and Amitabh Chandra, "Medicare Spending, The Physician Workforce, and The Quality Of Health Care Received By Medicare Beneficiaries." *Health Affairs*, April 2004: 184-97.
- Barnett, Henry, "Is Carotid Endarterectomy Safe in Patients over 80 Years Old?" *Nature Clinical Practice Cardiovascular Medicine* 2, 2005, 382-383.
- Berenson, Alex, "Hope, at \$4,200 a Dose," *The New York Times*, October 1, 2006: Section 3, P. 1, 7.
- Chan, An-Wen, Asbjorn Hrobjartsson, Mette T. Haahr, Peter C. Gotsche, and Douglas G. Altman, "Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles," *JAMA* 291(20), May 26, 2004: 2457-65.
- Chandra, Amitabh and Staiger, Douglas, "Testing a Roy Model with Productivity Spillovers: Evidence from the Treatment of Heart Attacks," *Journal of Political Economy* February 2007.
- CMS: Centers for Medicare and Medicaid Services, "Medicare Begins Performance-Based Payments for Physician Groups," Press Release, January 31, 2005, <http://www.cms.hhs.gov/apps/media/press/release.asp?Counter=1341>.
- Cortese, Denis and Robert Smoldt. Taking Steps Toward Integration. *Health Affairs*, January/February 2007; 26(1): w68-w71.
- Cutler, David M. and Mark McClellan, 2001. "Is Technological Change in Medicine Worth It?" *Health Affairs* (Sept/Oct): 11-29.
- Emanuel, Ezekiel, and Victor R. Fuchs, "Solved! It covers everyone. It cuts costs. It can get through Congress. Why Universal Healthcare Vouchers is the next big idea," *Washington Monthly*, June, 2005.
- Evans, DB, Tandon, A, Murray, CJ, and Lauer, JA. 2001 "Comparative Efficiency of National Health Systems: Cross-National Econometric Analysis," *BMJ* 323: 307-10.
- Fisher, Elliott, et al., "The Implications Of Regional Variations In Medicare Spending. Part 1: The Content, Quality, and Accessibility Of Care." *Annals of Internal Medicine*, 138(4), February 18, 2003: 273-87.
- Fisher, Elliott, et al., "The implications of regional variations in Medicare spending. Part 2: Health outcomes and satisfaction with care." *Annals of Internal Medicine*, 138(4), February 18, 2003,: 288-98.
- Fisher, Elliott, David E. Wennberg, Thérèse A. Stukel, and Daniel J. Gottlieb. Variations In The Longitudinal Efficiency Of Academic Medical Centers. *Health Affairs* Web Exclusive, October 7, 2004
- Fisher, Elliott S., Douglas O. Staiger, Julie P.W. Bynum, and Daniel J. Gottlieb, "Creating Accountable Care Organizations: The Extended Medical Staff," *Health Affairs*, December

2006:W44-W57.

Fuhrmans, Vanessa, "A Novel Plan Helps Hospital Wean Itself Off Pricey Tests," *Wall Street Journal*, January 12, 2007.

Garber, Alan M., "Cost-Effectiveness and Evidence Evaluation as Criteria for Coverage Policy," *Health Affairs* W4, May 2004: 284-96.

Goodman, David, Therese A. Stukel, Chiang-hua Chang, and John E. Wennberg, "End-of-life Care at Academic Medical Centers: Implications for Future Workforce Requirements," *Health Affairs*, 25(2), March/April 2006: 521-31.

Hannan, E.L., et al., "Improving the Outcomes of Coronary Artery Bypass Surgery in New York State," *JAMA* 271(10), 1994: 761-766.

Hayward, Rodney A., David M. Kent, Sandeep Vijan, and Timothy P. Hofer, "Reporting Clinical Trial Results to Inform Providers, Payers, and Consumers," *Health Affairs* 24(6), November/December 2005: 1571-81.

Hochman, Judith S., Gervasio A. Lamas, Christopher E. Buller, et. al., "Coronary Intervention for Persistent Occlusion after Myocardial Infarction," *New England Journal of Medicine*, 355(23), December 7, 2006: 2395-2407

Jencks, Stephen F., Edwin D. Huff, and Timothy Cuerdon, "Change in the Quality of Care Delivered to Medicare Beneficiaries, 1998-99 to 2000-2001." *JAMA* 289, January 15, 2003: 305-312.

Keeley EC, Boura JA, Grines CL. "Primary angioplasty versus intravenous thrombolytic therapy for acute myocardial infarction: a quantitative review of 23 randomised trials," *Lancet* 361, 2003: 13-20.

Kotlikoff, Laurence J. and Hagist, Christian, "Who's Going Broke? Comparing Growth in Healthcare Costs in Ten OECD Countries" (December 2005). NBER Working Paper No. W11833 Available at SSRN: <http://ssrn.com/abstract=875666>

Krumholz, Harlan M, Sharon-Lise T. Normand, John A. Spertus, David M. Shahlan, and Elizabeth H. Bradley, "Measuring Performance for Treating Health Attacks and Heart Failure: The Case for Outcomes Measurement," *Health Affairs* 26(1), January/February 2007: 75-85.

Jonathan Oberlander, "Health Reform Interrupted: The Unraveling Of The Oregon Health Plan," *Health Affairs*, 26(1), 2007: w96-w105.

Nord, Erik, Jeff Richardson, Andrew Street, Helga Kuhse, Peter Singer, "Who Cares about Cost? Does Economic Analysis Impose or Reflect Social Values?" *Health Policy*, 34(2), November 1995: 79-94.

O'Connor, G.T., et al., "A Regional Intervention to Improve the Hospital Mortality Associated with Coronary Artery Bypass Graft Surgery: The Northern New England Cardiovascular Disease Study Group," *JAMA* 275(11), 1996: 841-846.

Pearson, Steven, "Health Technology Assessment and Comparative Effectiveness: Recommendations for Improving Health Care Value in the United States," 2007.

Physicians' Working Group for Single-Payer National Health Insurance (PWG), "Proposal of the Physicians' Working Group for Single-Payer National Health Insurance," *JAMA* 290, 2003: 798-805.

Redford, Rita F., "Evidence, Appropriateness, and Technology Assessment in Cardiology: A Case Study of Computed Tomography," *Health Affairs*, January/February 2007: 86-95.

Second International Study of Infarct Survival (ISIS-2), "Randomized Trial of Intravenous Streptokinase, oral aspirin, Both or Neither Among 17,187 Cases of Suspected Acute Myocardial Infarction: ISIS-2," *Lancet* (1988), 2:349-360.

Skinner, Jonathan, Douglas Staiger, and Elliott Fisher, 2005. "Is Medical Technology Always Worth It? The Case of Acute Myocardial Infarction." *Health Affairs*, March/April 2006.

Skinner, Jonathan, and Douglas Staiger, "Technological Diffusion from Hybrid Corn to Beta Blockers", in E. Berndt and C. M. Hulten (eds.) *Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches*. University of Chicago Press and NBER (forthcoming). Also available as NBER working paper No. 11251 (April 2005).

Stukel, Therese, Lee Lucas, and David Wennberg, "Long-term Outcomes of Regional Variations in Intensity of Invasive vs Medical Management of Medicare Patients With Acute Myocardial Infarction" *JAMA* 293(11), March 16, 2005:1329-1337.

Yusuf, S., R. Peto, J. Lewis, R. Collins, and P. Sleight, "Beta Blockage During and After Myocardial Infarction: An Overview of the Randomized Trials." *Progress in Cardiovascular Disease* 27 (March/April, 1985): 335-71.

Weinstein, Milton, "From cost-effectiveness ratios to resource allocation: Where to draw the line?", in Frank Sloan F, ed., *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies*. New York: Cambridge University Press; 1996:77-97.

Welch, W.P., "Prospective Payment to Medical Staffs: A Proposal," *Health Affairs* 8(1), 1989: 34-49.

Welch, W.P., and M.E. Miller, "Proposals to Control High-Cost Hospital Medical Staffs," *Health Affairs* 13(4), 1994: 42-57.

Wennberg, David E., F.L. Lucas, John D. Birkmeyer, Carl E. Bredenberg, Elliott S. Fisher, "Variations in Carotid Endarterectomy Mortality in the Medicare Population: Trial Hospitals, Volume, and Patient Characteristics," *JAMA* 279(16), April 22/29, 1998: 1278-81.

Wennberg, John E., Elliott S. Fisher, and Jonathan Skinner, "Geography and the Debate over Medicare Reform," *Health Affairs* 2002: W96-W117.

Werner, Rachel M., Eric T. Bradlow. Relationship Between Medicare's Hospital Compare Performance Measures and Mortality Rates. *JAMA* 2006 296: 2694-2702.

Wilensky, Gail, Nicholas Wolter, and Michelle M. Fischer. Gain Sharing: A Good Concept Getting A Bad Name? *Health Affairs*, January/February 2007; 26(1): w58-w67.

Related Prior Publications

Geographic Variation In The Appropriate Use Of Cesarean Delivery

Do higher usage rates reflect medically inappropriate use of this procedure?

by Katherine Baicker, Kasey S. Buckles, and Amitabh Chandra

ABSTRACT: There is enormous geographic variation in the use of cesarean delivery: For births over 2,500 grams, adjusted cesarean rates vary fourfold between low- and high-use areas. Even for births under 2,500 grams, high-use counties have rates that are double those of low-use ones. Higher cesarean rates are only partially explained by patient characteristics but are greatly influenced by nonmedical factors such as provider density, the capacity of the local health care system, and malpractice pressure. Areas with higher usage rates perform the intervention in medically less appropriate populations—that is, relatively healthier births—and do not see improvements in maternal or neonatal mortality. [*Health Affairs* 25 (2006): w355–w367; 10.1377/hlthaff.25.w355]

THE CESAREAN SECTION RATE IN THE UNITED STATES is much higher than that in other countries.¹ Even within the United States, taking patients' risk factors into account, some areas use cesareans at much higher rates than others do: In 1996–2000, in Phoenix, Arizona, there were only fifteen cesareans per hundred births over 2,500 grams, while in Long Island, New York, there were twenty-six per hundred births. This is a source of national concern: The Centers for Disease Control and Prevention's (CDC's) Healthy People 2010 initiative has the explicit goal of reducing the cesarean birth rate.² This objective is predicated on the belief that high rates of cesarean delivery reflect procedure use in mothers and infants who obtain little benefit from the procedure. In the extreme, higher procedure rates might even be associated with iatrogenic harm, stemming from surgical complications that are not offset by therapeutic benefit.³

Uncovering the relationship between areawide intensity of use and the medical

Katherine Baicker is an associate professor in the Department of Public Policy, School of Public Affairs, at the University of California, Los Angeles. Kasey Buckles is an assistant professor of economics at the University of Notre Dame, in Notre Dame, Indiana. Amitabh Chandra (Amitabh_Chandra@harvard.edu) is an assistant professor of public policy in the John F. Kennedy School of Government, Harvard University, in Cambridge, Massachusetts. Baicker and Chandra are also affiliated with Dartmouth Medical School in Hanover, New Hampshire, and are faculty research fellows at the National Bureau of Economic Research in Cambridge, Massachusetts.

appropriateness of care requires a reliable way to measure appropriateness. There are major methodological challenges in using ex-post evaluation to determine the pervasiveness of inappropriate care. Typically, a systematic review of the literature is combined with the opinions of an expert clinical panel to score patients on a scale that measures appropriateness for a given procedure.⁴ Several studies have found that patients with the highest appropriateness scores benefit most from the intervention, and one notes that the views of practicing physicians are similar to those of an expert panel.⁵ Critics of this approach, however, note that often the variables used to define *appropriateness* have not been validated.⁶ In addition, criteria developed by different expert panels have been shown to exhibit enormous variability, particularly for procedure use classified as inappropriate, and to be greatly influenced by the composition of the panel.⁷

We introduce a new methodology to determine whether higher cesarean rates reflect less medically appropriate use of the procedure. We use the correlation between patient characteristics and whether or not that patient receives a cesarean section to construct a predicted probability of cesarean birth (PPC). For each birth, this is the probability that the typical obstetrician would perform a cesarean delivery, based on the patient's prebirth characteristics, and removing the effect of area characteristics that are unchanging over time. This measure has strengths and weaknesses: It can tell us which births are collectively viewed by doctors as being better candidates for a cesarean, but it cannot be used to infer the cutoff for a medically "appropriate" versus an "inappropriate" cesarean. Like all measures of medical appropriateness, it cannot inform us about whether the choice of cesarean delivery is associated with nonmedical factors such as patients' preferences; *appropriateness* in our paper refers only to medical appropriateness. Our index is not, however, subject to the arbitrary nature of measures calculated by panels or to the biases inherent in retrospective classification based on procedure outcomes. Nor, as we demonstrate below, is it confounded by characteristics of the area that are fixed over time. It is easily implemented and can be applied to any medical procedure.

Our analysis contains three parts. First, we demonstrate geographic variation in the use of cesarean delivery across different U.S. cities. We examined the correlates of higher use by studying the relationship between cesarean rates and birth characteristics, county socioeconomic characteristics, local provider capacity, and medical malpractice liability.⁸ The larger the role played by nonclinical factors such as provider supply and malpractice liability, the more likely it is that the marginal cesarean birth occurs for less medically driven reasons.

Second, we test the hypothesis that areas with higher cesarean rates are performing the intervention in births that are less medically appropriate for the use of cesarean delivery. If our hypothesis is correct, then the typical cesarean birth in more aggressive areas will have a lower PPC than the marginal cesarean birth in less aggressive areas, because physicians would have worked their way down the

distribution of patient appropriateness. Hence, this hypothesis implies a negative relationship between area cesarean rates and the average appropriateness of use of the procedure.

Third, we explore the hypothesis that even though areas with high cesarean rates might perform the procedure in less appropriate births, they might be better at the use of this procedure and achieve improved outcomes. To examine this theory, we studied the relationship between area cesarean rates and neonatal and maternal mortality ratios.

Study Data And Methods

■ **Data.** We used the National Center for Health Statistics (NCHS) linked birth and infant death data, pooled across the years 1995–1998, to calculate risk-adjusted county cesarean rates and the PPC (N = 15,592,980). We used these years of the birth data because we were able to obtain concurrent data on maternal mortality, medical malpractice, and other county characteristics. We selected the 10.2 million births that occurred during this period in the 198 U.S. counties with populations greater than 250,000 in the 1990 census because county of birth is not identified on birth certificates in smaller counties. Even if it were, sampling errors would make it difficult to calculate maternal and neonatal mortality for smaller geographic locales. Washington, D.C., was dropped from the analysis because medical malpractice data were not available for this area.

We classified infants as low birthweight (LBW) if their birthweight was less than 2,500 grams (n = 800,109) and as normal birthweight (NBW) otherwise (n = 9,361,844). Even with the large sample sizes at our disposal, we were unable to reliably estimate risk-adjusted cesarean rates separately for very-low-birthweight (VLBW) infants (those less than 1,500 grams, n = 151,480) in smaller cities. We therefore combined this group with LBW infants and use the term “LBW” in the text to designate all babies born under 2,500 grams. We defined *cesarean deliveries* to include both primary (n = 1,337,130) and repeat cesarean (n = 757,657). Using this definition, 20.6 percent of our sample had a cesarean delivery. This reflects a rate of 37.3 percent for LBW births and a rate of 19.2 percent for NBW births.⁹ In two unreported secondary analyses, we repeated our analysis using (1) a subsample of our data where repeat cesarean and vaginal birth after cesarean (VBAC) deliveries were excluded from the analysis, and (2) a sample of first births (where the decision to perform a repeat cesarean is not possible). Both secondary analyses yielded results that were virtually identical to the full-sample results reported below.

Data on physician and hospital resources, including the total number of physicians, pediatricians, and obstetrician-gynecologists (OB-GYNs) per birth; neonatal intensive care unit (NICU) beds per birth; and Medicaid share of inpatient days, are from the Area Resource Files (ARF). We also obtained data on county characteristics such as population, per capita income, urban classification, and demographic composition from the ARF.

We measured malpractice liability pressures in two different ways. First, we included the number and size of malpractice payments (arising from judgments and settlements and measured separately for surgery, obstetrics, and internal medicine) per physician in each state, ascertained from the National Practitioner Data Bank (NPDB). Our choice of these measures was motivated by research finding that physicians respond to both the number of claims and the average size of malpractice awards: Being sued imposes costs on physicians, including lost time at work and psychic costs.¹⁰ Ideally, we would have used claims per physician, but there is no nationally representative source for these data. This limitation would cause us to understate the potential role of malpractice liability. Second, we constructed a measure of malpractice liability pressure based on average physician malpractice premiums, as reported in the *Medical Liability Monitor*, a national survey of insurers. This measure addresses concerns that some payments may be missed by the NPDB and the fact that payments reported to the NPDB reflect claims filed a few years ago. A further advantage of using malpractice premiums as a measure of malpractice liability is that it reflects insurers' estimates of open and future claims—a factor that will be missed by the NPDB but might still affect physicians' practice style. Our measures are not mechanically correlated; indeed, other research has argued that a number of factors, including the interaction of state regulatory oversight and the insurance underwriting cycle, determine premiums.¹¹

County-level maternal mortality was calculated from the NCHS's multiple-cause-of-death mortality data from 1995 to 1998.¹² We counted any woman in an area ages 10–54 for whom “complications of pregnancy, childbirth, and the puerperium” is listed as the primary cause of death. We calculated maternal mortality ratios by dividing the count of maternal deaths by the number of live births that occurred in each county during the same time period.

■ **Analyses.** *Correlates of area-level cesarean rates.* We calculated unadjusted county-level cesarean rates and evaluated the correlates of this county-level intensity. We report analysis of variance (ANOVA) results, which explain the variance of these rates using four sets of covariates in a prespecified order that allows birth and socioeconomic status (SES) characteristics to have maximum explanatory power; doing so minimizes the role of nonmedical factors such as provider capacity and malpractice pressure: (1) an index of patient-level characteristics (calculated using predictions from a regression model of cesarean delivery on variables in the birth certificate data); (2) county-level measures of socioeconomic factors (including average income, unemployment rate, percentage living in poverty, percentage urban, percentage white, percentage of the population with less than a high school degree, high school and college degrees, percentage of hospital patient days eligible for Medicaid, and size of the population); (3) county-level provider characteristics (including total physicians, surgeons, pediatricians, OB-GYNs, internists, and other specialists per birth, as well as neonatal intensive care beds per birth); and (4) state-level medical malpractice liability (including the number and size of judgments and settle-

ments by medical specialty and malpractice premiums by specialty).¹³

Area usage rates and the predicted probability of cesarean birth. The second step in our analysis was to correlate risk-adjusted area-level variations in cesarean usage rates with our measure of the average appropriateness of patients who received the procedure. We hypothesized that in areas with higher cesarean rates, the intervention is being used for births that are less medically appropriate for it. If this were true, we should observe a negative relationship between the PPC for all births by cesarean and area cesarean rates. Note that this is not a mechanical relationship—both a positive relationship and no relationship are also possible. The former would occur if physicians are first performing cesareans based on some unobservable characteristic (for example, patients' income or insurance coverage), regardless of medical appropriateness, before moving on to more appropriate patients. The latter would occur if some areas are uniformly more aggressive and perform cesareans without regard to clinical appropriateness. No relationship would also be observed if half of physicians performed cesareans by the first effect and half by the second.

We computed risk-adjusted county cesarean rates by estimating a regression model of the probability that an individual infant is delivered via cesarean on patient-level covariates and indicator variables for each of the identified counties in our data. The coefficients on the county indicator variables estimated the risk-adjusted probability of receiving a cesarean delivery in each particular county.¹⁴

The PPC was also obtained from this regression, but it relies on only the portion of the prediction that relies on birth certificate data; it excludes the county fixed effect.¹⁵ Thus, the PPC measures which births are more likely to involve a cesarean, based on clinical characteristics but independent of the local practice style. The regressions that generated these rates yielded an R^2 of 0.39 for NBW babies and 0.32 for LBW babies. Thus, we are able to explain at least as much of the variation in cesareans as previous research, which finds an R^2 of 0.37 using both birth certificate data and discharge data to construct risk-adjusted cesarean rates.¹⁶ We report least squares regressions (WLS) to examine the relationship between the risk-adjusted cesarean rate for each birthweight category and the PPC at the county level.

Procedure intensity and mortality. The third step in our analysis was to evaluate the relationship between (1) variations in the intensity of the use of cesareans and (2) maternal and infant mortality across areas. Even if physicians in areas that use cesareans more intensively are performing the procedure on less and less appropriate patients, their patients could still have better outcomes if the physicians are more skilled at the procedure. We tested this hypothesis by regressing both risk-adjusted infant mortality and maternal mortality on risk-adjusted cesarean rates. We performed this analysis using both WLS and negative binomial regression and obtained quantitatively similar estimates. To preserve transparency, we report the results from the former technique.

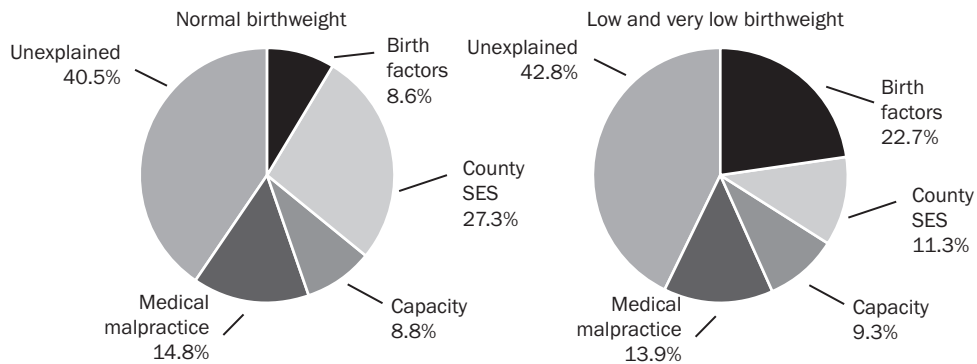
Study Results

■ **Correlates of area-level variation in the use of cesareans.** Exhibit 1 illustrates the variation in unadjusted cesarean rates that can be explained by patient and area characteristics for births over 2,500 grams and those under 2,500 grams. Characteristics of the birth explain much more of the variation in cesarean rates for LBW infants than they do for NBW infants. For both NBW and LBW births, a large part of the variation remains unexplained.

Exhibit 2 reports the range of unadjusted and risk-adjusted cesarean rates for the United States and the fifty-one counties with the most births between 1995 and 1998.¹⁷ With few exceptions, risk adjustment does not greatly alter the ranking of the counties, and the correlation coefficients between area cesarean rates obtained from unadjusted and risk-adjusted models are 0.93 ($p < .001$) for NBW births and 0.84 ($p < .001$) for LBW births.¹⁸ Counties whose risk-adjusted rates are higher than the unadjusted rates have cesarean usage that is more intensive than what would be predicted using observable characteristics. The variation in the rates is large: Even in the largest U.S. counties, risk-adjusted rates ranged from 6.7 percent to 28.9 percent for NBW births and from 25 percent to 50 percent for LBW births.

■ **Cesarean delivery and medical appropriateness.** In Exhibit 3 the average PPC of patients receiving a cesarean in each area is plotted against risk-adjusted area-level cesarean rates. LBW babies are more likely to be born via cesarean than

EXHIBIT 1
Correlates Of County Cesarean Rates For Births With Normal (2,500 Grams Or More) And Low And Very Low (Less Than 2,500 Grams) Birthweight, 1995–1998



SOURCE: Authors' calculations from U.S. natality data, 1995–1998.

NOTES: Fraction of variation in use of cesareans across 198 U.S. counties in 1995–1998 attributable to different correlates, including patient-level characteristics (mother's age, race, education, and marital status; adequacy of prenatal care; medical risk factors; congenital anomalies; and complications of labor and delivery), county-level measures of socioeconomic status (SES) factors (average per capita income; unemployment rate; percentage living in poverty; percentage urban; percentage minority; percentage with less than high school, high school diploma, and college degree; percentage of hospital patient days eligible for Medicaid; and population), county-level provider supply (total physicians, surgeons, pediatricians, obstetrician-gynecologists, internists, and other specialists per birth, as well as the number of neonatal intensive care unit beds per birth), and state-level malpractice liability (the number and size of judgments and settlements by specialty and malpractice premiums by specialty).

EXHIBIT 2
Cesarean Rates For Normal-Birthweight And Low/Very-Low-Birthweight Births For
The Fifty-One U.S. Counties With The Largest Number Of Births, 1995–1998

County	State	Major city	Unadjusted rates (%)		Adjusted rates (%)	
			Normal birthweight	Low/very low birthweight	Normal birthweight	Low/very low birthweight
Hennepin	MN	Minneapolis	13.4	37.0	12.5	35.4
Milwaukee	WI	Milwaukee	13.5	26.7	13.5	29.8
Denver	CO	Denver	13.7	35.0	15.8	32.0
Salt Lake	UT	Salt Lake City	14.4	35.6	15.8	32.2
Maricopa	AZ	Phoenix	14.7	34.8	17.6	35.0
Franklin	OH	Columbus	15.6	36.1	18.0	35.9
Alameda	CA	Oakland	15.9	33.2	19.2	38.5
Travis	TX	Austin	16.4	35.8	17.3	36.7
Santa Clara	CA	Santa Clara	16.6	37.3	18.0	36.1
Wayne	MI	Detroit	16.7	31.1	18.5	35.2
King	WA	Seattle	16.8	36.3	14.0	31.5
Sacramento	CA	Sacramento	17.0	38.1	18.0	37.9
Cook	IL	Chicago	17.0	32.3	19.2	35.7
Baltimore City	MD	Baltimore	17.5	29.4	19.5	35.2
Hamilton	OH	Cincinnati	17.6	38.3	17.2	35.0
Bronx	NY	Bronx	17.9	31.5	19.8	36.1
Allegheny	PA	Pittsburgh	17.9	34.0	16.2	31.2
Fulton	GA	Atlanta	18.1	34.2	19.4	38.9
Orange	FL	Orlando	18.2	37.1	20.1	37.4
Cuyahoga	OH	Cleveland	18.2	34.1	17.8	34.5
Philadelphia	PA	Philadelphia	18.4	31.6	16.3	35.2
Marion	IN	Indianapolis	18.5	36.3	19.3	36.5
Clark	NV	Las Vegas	18.7	38.8	20.9	41.3
Orange	CA	Long Beach/Santa Ana	18.9	40.7	21.5	39.8
Riverside	CA	Riverside	18.9	39.1	23.7	44.2
El Paso	TX	El Paso	19.0	36.8	19.7	37.3
Middlesex	MA	Cambridge/Boston	19.0	33.7	15.0	33.6
San Diego	CA	San Diego	19.3	38.9	21.9	42.0
Suffolk	MA	Boston	19.4	42.5	15.5	32.9
Tarrant	TX	Fort Worth	19.6	40.8	20.3	40.9
Montgomery	MD	Greater DC area	19.7	38.8	17.5	35.8
San Bernardino	CA	San Bernardino	19.8	41.4	23.3	43.5
Dallas	TX	Dallas	20.0	36.2	20.2	36.3
Bexar	TX	San Antonio	20.2	39.6	20.2	39.6
St. Louis	MO	St. Louis	20.3	39.0	17.0	33.8
Kings	NY	Brooklyn	20.3	34.1	21.1	38.1
Oakland	MI	Greater Detroit	20.4	39.6	17.7	35.6
Fairfax	VA	Fairfax	20.5	42.5	19.6	40.3
Shelby	TN	Memphis	21.1	33.8	19.9	37.0
Palm Beach	FL	Palm Beach	21.1	37.1	21.8	38.2
Harris	TX	Houston	21.1	38.5	21.1	38.8
Fresno	CA	Fresno	21.1	37.5	26.4	42.5
New York	NY	Manhattan	21.3	41.7	18.4	39.4
Queens	NY	Queens	21.9	36.7	21.6	39.5
Los Angeles	CA	Los Angeles	22.0	42.5	25.0	45.9
Essex	NJ	Newark	23.6	39.9	20.5	38.2
Hillsborough	FL	Tampa	23.7	43.9	24.3	44.1
Nassau	NY	Greater New York City	23.9	45.3	21.5	40.9
Miami-Dade	FL	Miami	24.5	36.8	25.6	41.8
Broward	FL	Ft. Lauderdale	24.5	43.8	24.4	44.8
Suffolk	NY	Long Island	26.0	45.1	21.3	39.9

EXHIBIT 2
Cesarean Rates For Normal-Birthweight And Low/Very-Low-Birthweight Births For The Fifty-One U.S. Counties With The Largest Number Of Births, 1995–1998 (cont.)

County	State	Major city	Unadjusted rates (%)		Adjusted rates (%)	
			Normal birthweight	Low/very low birthweight	Normal birthweight	Low/very low birthweight
U.S. rate ^a			19.2	37.3	19.2	37.3
Range: 1st–99th percentile			11.9–28.7	24.4–49.9	6.7–28.9	25.0–50.0

SOURCE: Authors' calculations from U.S. natality data, 1995–1998.

NOTES: Risk-adjusted cesarean rates for normal-birthweight (NBW) and low-birthweight (LBW)/very-low-birthweight (VLBW) births in the fifty-one U.S. counties with the most births in 1995–1998 (n = 10,161,953). Counties are listed in descending order of the total number of births. The correlation between NBW and LBW/VLBW cesarean rates is 0.87 (p < .001), and the correlation between adjusted and unadjusted rates is 0.71 (p < .001).

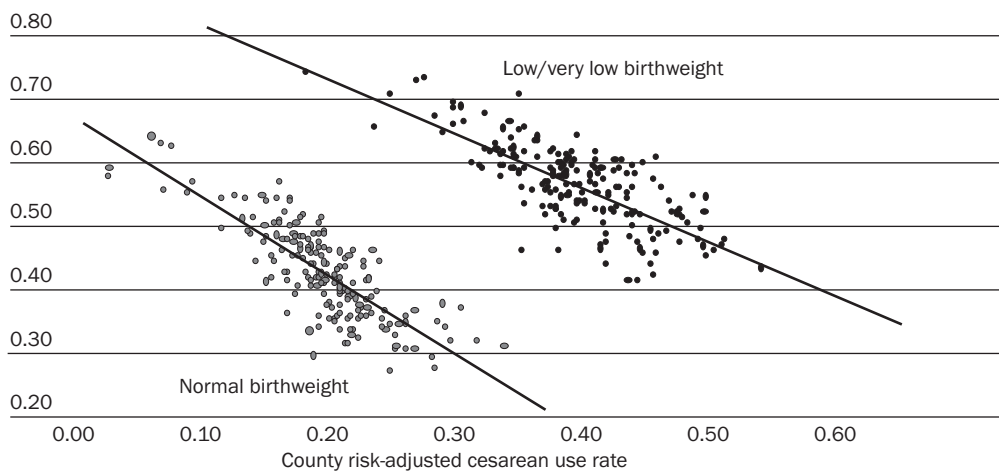
^aThe 198 largest U.S. counties.

NBW babies, but even for these births, the variation in risk-adjusted rates across large counties is striking. For both NBW and LBW babies, the average birth through cesarean has a systematically lower PPC in areas that do more cesarean sections. Increases in the cesarean rate are associated with major declines in the average medical appropriateness of births.

■ **Cesarean use and patient mortality.** Our third analytic step was to examine

EXHIBIT 3
Relationship Between Predicted Probability Of Cesarean Birth And Cesarean Rates, Normal Birthweight And Low/Very Low Birthweight, 1995–1998

Predicted probability for cesarean (PPC) for all cesarean births



SOURCE: Authors' calculations from U.S. natality data, 1995–1998.

NOTES: Relationship between each area's cesarean rate and the average probability of cesarean delivery among cesarean births in the area (a measure of appropriateness). A ten-percentage-point increase in the cesarean rate for normal-birthweight births (2,500 grams or higher) leads to a fourteen-percentage-point reduction in appropriateness for cesarean births (p < .001) and a nine-percentage-point decline in appropriateness for low-birthweight and very-low-birthweight births (less than 2,500 grams) (p < .001). Observations are 198 U.S. counties (1995–1998).

the effect of risk-adjusted area cesarean rates on risk-adjusted infant and maternal mortality rates (Exhibit 4). For NBW births, the standard deviation of the cesarean rate was 3.2 percentage points, and a change in the cesarean rate of this magnitude (from the average rate of 19.2 percent) would not alter neonatal mortality. Similarly, for LBW births, the standard deviation of the cesarean rate was 4.8 percentage points. Altering the cesarean rate from 37.5 percent upward by this amount would also not result in a change in neonatal mortality. The exhibit also shows the effect of changes in use of cesareans on maternal mortality. There is a small but statistically insignificant relationship: The average maternal mortality ratio is 7.3 deaths per 100,000 births, and a 3.2-percentage-point decrease in the cesarean rate is associated with an insignificant change in the maternal mortality rate ($p = .06$).

Discussion

We have demonstrated that there is large geographic variation in the use of cesarean delivery, only some of which is explained by patients' characteristics and county SES measures. A substantial portion of this variation remains unexplained. This unexplained variation could be labeled as the "practice style" of an area, if it is unrelated to patient and area characteristics. In fact, if the number of physicians or NICU beds in an area is also a consequence of underlying variations in practice style, then we might have understated the role of local practice style. On the other hand, if county cesarean rates are partially determined by patients' preferences, then we have overstated the case for physician practice style. To the extent that patients' preferences are explained by the demographic variables that are controlled for in our analysis (age; race; maternal schooling; and county SES measures such as income, poverty, and population), our labeling of the unexplained varia-

EXHIBIT 4 Association Between Higher Cesarean Use Rates And Neonatal And Maternal Mortality, 1995–1998

	Neonatal mortality		Maternal mortality
	Normal birthweight	Low and very low birthweight	All birthweights
Cesarean use rate	19.2%	37.5%	20.5%
Standard deviation (SD) of cesarean use rate ^a	3.2	4.8	3.2
Mortality rate per 10,000 births	9.7	330.4	0.73
Effect of decreasing cesarean rate by 1 SD on mortality per 10,000 births	0.2 ($p < .97$)	0.2 ($p < .28$)	0.096 ($p < .10$)

SOURCE: Authors' calculations from U.S. natality data, 1995–1998.

NOTES: Estimated effect of differences in cesarean rates on infant and maternal mortality. Relationships were estimated using births in 198 U.S. counties, 1995–1998. Decreasing the use of cesareans by 4.8 percentage points (one standard deviation) among low-birthweight (LBW)/very-low-birthweight (VLBW) births would decrease neonatal mortality by 0.2 per 10,000 births, but this change is not statistically significant.

^a Percentage points.

tion as “physician practice style” is justified.

We demonstrate that more-aggressive areas perform the procedure for births that are less medically appropriate for the procedure. This finding is a prerequisite for demonstrating that the higher rates are symptomatic of “flat-of-the-curve” medicine, where physicians work into less appropriate populations. Indeed, our analysis challenges previous work that found no relationship between the intensity of diagnostic testing and clinical indications for the use of that test.¹⁹ We demonstrate that some physicians are not systematically more aggressive than others (that is, do not have a disposition to be more aggressive on all births regardless of medical appropriateness); rather, physicians rank patients on a distribution of clinical appropriateness and work their way down that distribution. The point at which they stop in that distribution is affected by nonmedical factors such as provider capacity, malpractice liability, and local physicians’ opinion.

■ **Study limitations.** Our analysis is not without limitations. We relied on birth certificate data; important information that is available to the physician, such as maternal drug use and detailed medical histories, was not available to us. We note that our flexible birth-level risk-adjustment models have explanatory power that is identical to other studies that used both birth certificate and hospital discharge data. Although this is reassuring, it does not mean that omitted factors are not an issue for our study. For omitted factors to bias our analysis (by overstating the role of physician practice style), they would have to be more prevalent in areas with higher cesarean rates and completely uncorrelated with the maternal characteristics and county socioeconomic factors that we controlled for. To explore this possibility, we restricted our sample of births to states that report information on tobacco use, alcohol consumption, and weight gain during pregnancy, and we included these variables as additional covariates in our analysis. The correlation between risk-adjusted area cesarean rates using these additional covariates and those reported in Exhibit 2 was 0.98 ($p < .001$) for LBW births and 0.99 ($p < .001$) for NBW births. These findings suggest that certain types of omitted clinical variables might not be of first-order concern.

■ **Policy implications.** Our finding that physicians in areas with higher cesarean rates are performing procedures that are of decreasing medical value to patients has important policy implications. Cesareans are an expensive intervention, with an average cost in 2003 of \$12,468—twice the cost of the average vaginal birth (\$6,240). There is also evidence that women undergoing a cesarean delivery are at much higher risk for rehospitalization for uterine infection and obstetrical surgical wound complications.²⁰ The real costs of a cesarean delivery might therefore be much higher than we have stated. Our analysis demonstrates that reductions in the cesarean rate in high-use counties (of the magnitude of three to five percentage points) will not affect mortality among newborns and mothers. Reductions in the cesarean rate have been demonstrated to be achievable in the clinical literature: A hospital was able to reduce its rate from 17.5 percent to 11 percent over two years without any

adverse health outcomes. The reduction was achieved by requiring a second opinion, by providing objective criteria for when a cesarean delivery is indicated, and by a review of all cesarean deliveries. If the cesarean rate is to be reduced, we would also argue that reductions should be targeted toward primary cesareans and not repeat cesareans. The latter are believed to be safer than the VBAC alternative, and each primary cesarean that is averted also eliminates the need for a repeat cesarean.²¹

Reductions in the cesarean rate could deleteriously affect other health outcomes that were not examined in our study.²² Furthermore, if county cesarean rates reflect patients' preferences for elective cesareans, then reducing the rate will reduce patient satisfaction. Although there has probably been an increase in patient demand for elective cesarean delivery, it is not known if this explains the geographic variation in cesarean rates, and it is difficult to rationalize preferences as being the explanation for the ten-percentage-point difference in cesarean rates between Minneapolis and Miami. The importance of preferences can be partially assessed by examining the role of county socioeconomic factors that ought to be correlated with patients' preferences (such as income, educational attainment, percentage minority, and percentage metropolitan). Relative to the unexplained variation and variation explained by medical malpractice and capacity factors, the role of county factors is small, and it suggests that preferences cannot be the principal driver of geographic variation in cesarean rates. Likewise, patient satisfaction has been shown to be unaffected by the presence of intensive health care.²³

Variation in local medical opinion about the right cutoff for initiating a cesarean delivery, strengthened by available capacity and malpractice pressure, continues to be the best explanation for the facts in our analysis. In an era of soaring health care costs, where already strained public programs reimburse for cesarean delivery, it seems particularly important to consider the ramifications of intensive treatments whose medical benefits are uncertain when performed in less medically appropriate populations.

.....
Katherine Baicker and Amtabh Chandra acknowledge funding from the National Institute on Aging, Grant no. NIA P01 AG19783-02; and Chandra, from the National Institute of Child Health and Human Development, Grant no. NICHD R01 HD44003-01. The authors thank three astute reviewers, Jonathan Skinner, Douglas Staiger, and Jack Wennberg, for influencing their thinking on this topic. The opinions in this paper are those of the authors and do not reflect those of any institution they are affiliated with or have received support from.

NOTES

1. J.M. Belizan et al., "Rates and Implications of Caesarean Sections in Latin America: Ecological Study," *British Medical Journal* 319, no. 7222 (1999): 1397–1400; and B.L. Flamm, "Caesarean Section: A Worldwide Epidemic?" *Birth* 27, no. 2 (2000): 139–140.
2. U.S. Department of Health and Human Services, *Healthy People 2010* (Washington: U.S. Government Printing Office, November 2000).
3. E.S. Fisher and H.G. Welch, "Avoiding the Unintended Consequences of Growth in Medical Care: How Might More Be Worse?" *Journal of the American Medical Association* 281, no. 5 (1999): 446–453.
4. M.R. Chassin et al., "Does Inappropriate Use Explain Geographic Variations in the Use of Health Care Services? A Study of Three Procedures," *Journal of the American Medical Association* 258, no. 18 (1987): 2533–2537; L.L. Leape et al., "Does Inappropriate Use Explain Small-Area Variations in the Use of Health Care Services?" *Journal of the American Medical Association* 263, no. 5 (1990): 669–672; C.M. Winslow et al., "The Appropriateness of Carotid Endarterectomy," *New England Journal of Medicine* 318, no. 12 (1988): 721–727; and C.M. Winslow et al., "The Appropriateness of Performing Coronary Artery Bypass Surgery," *Journal of the American Medical Association* 260, no. 4 (1988): 505–509.
5. See, for example, Chassin et al., "Does Inappropriate Use Explain Geographic Variations?"; Winslow, "The Appropriateness of Carotid Endarterectomy"; H. Hemingway et al., "Underuse of Coronary Revascularization Procedures in Patients Considered Appropriate Candidates for Revascularization," *New England Journal of Medicine* 344, no. 9 (2001): 645–654; and J.Z. Ayanian et al., "Rating the Appropriateness of Coronary Angiography—Do Practicing Physicians Agree with an Expert Panel and with Each Other?" *New England Journal of Medicine* 338, no. 26 (1998): 1896–1904.
6. J.P. Kassirer, "The Quality of Care and the Quality of Measuring It," *New England Journal of Medicine* 329, no. 17 (1993): 1263–1265; N.R. Hicks, "Some Observations on Attempts to Measure Appropriateness of Care," *British Medical Journal* 309, no. 6956 (1994): 730–733; and C.E. Phelps "The Methodologic Foundations of Studies of the Appropriateness of Medical Care," *New England Journal of Medicine* 329, no. 17 (1993): 1241–1245.
7. P.G. Shekelle et al., "The Reproducibility of a Method to Identify the Overuse and Underuse of Medical Procedures," *New England Journal of Medicine* 338, no. 26 (1998): 1888–1895.
8. We selected these factors based on variables suggested by previous research. See, for example, A.A. Kabir et al., "Unnecessary Cesarean Delivery in Louisiana: An Analysis of Birth Certificate Data," *American Journal of Obstetrics and Gynecology* 190, no. 1 (2004): 10–19; D.C. Aron et al., "Variations in Risk-Adjusted Cesarean Delivery Rates According to Race and Health Insurance," *Medical Care* 38, no. 1 (2000): 35–44; A.A. Kabir et al., "Racial Differences in Cesareans: An Analysis of US. 2001 National Inpatient Sample Data," *Obstetrics and Gynecology* 105, no. 4 (2005): 710–718; W.J. Hueston and S. Lewis-Stevenson, "Provider Distribution and Variations in Statewide Cesarean Section Rates," *Journal of Community Health* 26, no. 1 (2001): 1–10; L. Dubay, R. Kaestner, and T. Waidmann, "The Impact of Malpractice Fears on Cesarean Section Rates," *Journal of Health Economics* 18, no. 4 (1999): 491–522; and L.R. Burns, S.E. Geller, and D.R. Wholey, "The Effect of Physician Factors on the Cesarean Section Decision," *Medical Care* 33, no. 4 (1995): 365–382.
9. Descriptive statistics for our sample are reported in Appendix Exhibit 1, available online at <http://content.healthaffairs.org/cgi/content/full/25/w355/DC2>.
10. D.P. Kessler and M.B. McClellan, "How Liability Law Affects Medical Productivity," *Journal of Health Economics* 21, no. 6 (2002): 931–955.
11. Congressional Budget Office, *Limiting Tort Liability for Medical Practice*, 8 January 2004, <http://www.cbo.gov/showdoc.cfm?index=4968&sequence=0> (accessed 5 July 2006); U.S. Government Accountability Office, *Medical Malpractice Insurance: Multiple Factors Have Contributed to Increased Premium Rates* (Washington: GAO, 2003); and K. Baicker and A. Chandra, "The Effect of Malpractice Liability on the Delivery of Health Care," *Forum for Health Economics and Policy*, Forum: Frontiers in Health Policy Research, vol. 8, article 4 (2005), <http://www.bepress.com/fhep/8/4> (accessed 21 May 2006).
12. Centers for Disease Control and Prevention, Compressed Mortality Database, 1995–1998, <http://wonder.cdc.gov/mortICD9J.html> (accessed 23 August 2005).
13. The patient-level controls include birthweight in grams, gestation in weeks, mother's age, mother's race (black, white, or other), mother's education (fewer than 9 years, 9–11 years, 12 years, 13–15 years, and 16 or more years), mother's marital status (married, not married), birth order (first, second, third, or fourth or higher), newborn's sex, prenatal care use based on a Kessner index (adequate, intermediate, or inadequate), and indicators for the presence of each medical risk factor, congenital anomaly, or complications of labor and delivery.

14. We also estimated more complicated models that are, in principle, more efficient. We estimated (1) a conditional logit model, (2) a random-intercept logistic regression, and (3) a generalized estimating equation (GEE) model with a logit link function and exchangeable correlation structure. The correlation coefficients between the county effects obtained from these models and our preferred model are 0.981 ($p < .001$), 0.95 ($p < .001$), and 0.962 ($p < .001$), respectively. Given the similarity of area rankings across these different procedures, we favored the more transparent model.
15. Formally, for birth i (with characteristics X_i) in county j (denoted by C_j), the propensity score for a cesarean section (CS) is: Probability $CS_{ij} = 1 = F(X_i b + C_j)$. $F(\cdot)$ is the cumulative distribution for the logistic density, and we estimate this model using fixed-effects regression (thereby relaxing the assumption of GEE or random effects models that C is not correlated with X). C_j represents the county risk-adjusted cesarean rate. The PPC is the predicted probability of receiving a birth based only on birth characteristics, $Pr(CS_{ij} = 1) = F(X_i b)$ averaged across all births in a county.
16. E.B. Keeler et al., "Adjusting Cesarean Delivery Rates for Case Mix," *Health Services Research* 32, no. 4 (1997): 511–528.
17. Appendix Exhibit 3 provides this information for all 198 counties in our study. See Note 9.
18. Rates for NBW and LBW births are highly correlated within counties (correlation coefficient = 0.87, $p < .001$), which allays concern that the variation in cesarean rates among LBW babies is driven by random "noise."
19. Chassin et al., "Does Inappropriate Use Explain Geographic Variations?"; and Leape et al., "Does Inappropriate Use Explain Small-Area Variations?"
20. M. Lydon-Rochelle et al., "Association between Method of Delivery and Maternal Rehospitalization," *Journal of the American Medical Association* 283, no. 18 (2000): 2411–2416.
21. M.B. Landon et al., "Maternal and Perinatal Outcomes Associated with a Trial of Labor after Prior Cesarean Delivery," *New England Journal of Medicine* 351, no. 25 (2004): 2581–2589.
22. In work not reported in this paper, we examined other outcomes such as birth injuries and complications of delivery. We did not find that higher (risk-adjusted) cesarean rates were associated with improvements in these outcomes but did not pursue this line of inquiry after noting the poor data quality of these fields in birth-certificate data. See, for example, S. Northam and T. Knapp, "The Reliability and Validity of Birth Certificates," *Journal of Obstetric, Gynecologic, and Neonatal Nursing* 35, no. 1 (2006): 3–12; and D.L. DiGiuseppe et al., "Reliability of Birth Certificate Data: A Multi-Hospital Comparison to Medical Records Information," *Maternal and Child Health Journal* 6, no. 3 (2002): 169–179.
23. E.S. Fisher et al., "The Implications of Regional Variations in Medicare Spending, Part 2: Health Outcomes and Satisfaction with Care," *Annals of Internal Medicine* 138, no. 4 (2003): 288–298; and K. Baicker and A. Chandra, "Medicare Spending, the Physician Workforce, and Beneficiaries' Quality of Care," *Health Affairs* 23 (2004): w184–w197 (published online 7 April 2004; 10.1377/hlthaff.23.w4.184).

Malpractice Liability Costs And The Practice Of Medicine In The Medicare Program

This analysis suggests that an important association exists between malpractice costs and the use of imaging services in particular.

by **Katherine Baicker, Elliott S. Fisher, and Amitabh Chandra**

ABSTRACT: Mounting malpractice liability costs might affect physician practice patterns in many ways, such as increasing the use of diagnostic procedures while reducing major surgeries. This paper quantifies the association between malpractice liability costs and the use of physician services in Medicare. We find that higher malpractice awards and premiums are associated with higher Medicare spending, especially for imaging services that are often believed to be driven by physicians' fears of malpractice. The 60 percent increase in malpractice premiums between 2000 and 2003 is associated with an increase in total Medicare spending of more than \$15 billion. [*Health Affairs* 26, no. 3 (2007): 841–852; 10.1377/hlthaff.26.3.841]

RECENT INCREASES IN PHYSICIAN MALPRACTICE PREMIUMS and rapid growth in the number and size of awards to plaintiffs have raised widespread concerns about the medical malpractice liability system.¹ Although some argue that the current system plays an important role in maintaining the quality of care, others point out that it fails to compensate most patients who suffer avoidable injuries and punishes many physicians for adverse events that were not caused by negligence.² Perhaps even greater concerns have been raised about how rising malpractice premiums and payments affect the way that medicine is practiced.³

We focus on state-level variation in malpractice costs and health care use and spending patterns in the Medicare population from 1993 to 2001. We hypothesize that the practice of medicine—and the use of physician services in particular—

Katherine Baicker is an associate professor of public policy in the School of Public Affairs, University of California, Los Angeles. Elliott Fisher is a professor at Dartmouth Medical School in Hanover, New Hampshire, and a member of the Veterans Affairs Outcomes Group, Veteran Affairs Medical Center, White River Junction, Vermont. Amitabh Chandra (Amitabh_Chandra@Harvard.edu) is an assistant professor of public policy, John F. Kennedy School of Government, Harvard University, in Cambridge, Massachusetts. All authors are associates of the Center for the Evaluative Clinical Sciences, Dartmouth Medical School. Baicker and Chandra are also faculty research fellows at the National Bureau of Economic Research in Cambridge, Massachusetts.

“We hypothesize that the effect of increasing liability will be most pronounced for common, discretionary physician services.”

.....

will be different in states in which physician malpractice liability costs are higher (as measured by higher premiums or malpractice payments).⁴

Previous research on the effect of malpractice costs on the practice of medicine has focused on the use of a relatively small set of specific procedures, physician surveys of “consciously defensive” medicine, or comparisons of hospital spending on heart attack patients in states with and without tort-reform initiatives.⁵ These analyses do not quantify the aggregate effect of an increase in malpractice liability on clinical practice, total spending, or spending on physician services. Furthermore, many of these studies were conducted prior to the mid-1990s. Since then, there have been major changes in medical technology, including the increased use of diagnostic imaging tests, medical management, and minimally invasive surgery.⁶

We hypothesize that the effect of increasing liability will be most pronounced for common, discretionary physician services (such as visits, consultations, diagnostic tests, imaging services, and minor procedures, where errors of omission are perceived to carry greater malpractice risks than errors of commission) or for discretionary procedures where physicians may decline treatment for risky patients altogether. The effect on total use is ambiguous. On the one hand, increased testing might lead to some additional downstream treatment as a result of the additional medical services required to treat conditions not identified in areas with lower testing rates.⁷ On the other hand, concerns about malpractice could lead to lower rates of elective surgery if physicians leave areas with unfavorable malpractice climates or seek to avoid some higher-risk procedures or patients.⁸ We therefore hypothesize that any effect on total spending will be smaller than the effect on low-risk discretionary physician services.

Health care spending that is induced by malpractice costs and that costs more than it benefits patients (through improvements in mortality, morbidity, and patient satisfaction) is often labeled “defensive medicine.”⁹ Our analysis speaks primarily to the changes in use that are associated with changes in malpractice costs, but we also provide some evidence on whether additional spending is associated with improvements in mortality.

Study Data And Methods

■ **Analysis.** We report regression-adjusted estimates of the association between the growth of state-level malpractice payments per physician (or malpractice premiums) and the growth of use of and spending on several different types of procedures between 1993 and 2001. We performed our analysis at the state level because many aspects of the medical liability and medical practice environment (such as tort reforms) are set at that level.¹⁰ We focused on changes in malpractice costs and

changes in health care spending within states, to account for any confounders that are time-invariant within each state. For example, if a certain state was more urban or had a more heavily regulated health care sector (which might influence both practice patterns and malpractice liability exposure) than others, the effect of that factor would be netted out of our longitudinal analysis. This longitudinal analysis also accounted for tort reforms that were implemented before 1993 or remained unchanged through 2001 (as the vast majority of reforms were).¹¹ Our choice of study periods was further motivated by data availability and by the fact that the long window reduces the effect of measurement error.¹²

To control for factors that vary over time at the state level and might be correlated with malpractice liability and medical care use rates, we included covariates for per capita income, unemployment rate, education levels, racial composition, hospital beds per capita, and health maintenance organization (HMO) penetration.¹³ To validate these results, we examined whether our measures of malpractice liability were associated with outcomes that were unlikely to be uninfluenced by that liability, such as hospitalizations for hip fracture and acute myocardial infarction. It is unlikely that the incidence of or hospitalizations for these diseases were driven by the malpractice environment, although they were likely affected by potential confounders such as the underlying health of the population, so estimating the effect of malpractice liability costs on these outcomes can help test our methodological design.

■ **Defining “malpractice liability.”** We constructed two independent measures of malpractice liability costs. Our primary measure was the mean dollar value of malpractice payments (arising from both judgments and settlements) per physician in each state. Our choice of this measure was motivated by research finding that physicians respond to the number of claims as well as to the average size of malpractice awards: Being sued imposes costs on physicians, including lost time at work and psychic costs.¹⁴

We constructed this measure using data from the National Practitioner Data Bank (NPDB).¹⁵ We examined payments that resulted from either a court judgment or a settlement made outside of the courts. We averaged data for each of two periods, 1991–1993 and 1999–2001. Although the number of claims per physician would provide an additional measure of the burden of malpractice liability on practicing physicians, no national data on claims were available, and studies suggest that payments per physician are highly correlated with claims per physician.¹⁶ Despite limitations (such as the “corporate shield” loophole and potential under-reporting), researchers report that the NPDB is the most representative national database on medical malpractice payments and that the size of these potential biases is limited.¹⁷

Nevertheless, to address concerns that some payments might be missed by the NPDB and that payments reported to the NPDB reflect claims filed a few years ago, we constructed a second measure of malpractice liability costs based on phy-

sician malpractice insurance premiums. A further advantage of using malpractice premiums as a measure of malpractice liability is that they reflect insurers' estimates of open and future claims—a factor that will be missed by the NPDB. Our measure was constructed from premiums reported in the *Medical Liability Monitor* (MLM), whose annual national survey of insurers provides premium data for internal medicine, general surgery, and obstetrics-gynecology by state. We calculated average premiums faced by a typical physician in a state by weighting premium data across specialties by the physician mix in each state and averaging three years of data to minimize idiosyncrasies. Our final data consist of average premiums by state for 1991–1993 and 2000–2002, adjusted for inflation.¹⁸

■ **Use of medical care.** Our measures of use and spending were both based on the Medicare population, chosen because of the unusually rich data available. Our primary dependent variables were (1) total Medicare spending per beneficiary, (2) spending per beneficiary on each major component of total professional and laboratory services (evaluation and management, diagnostic tests, imaging, minor procedures, major procedures) based on the Berenson-Eggers Type of Service (BETOS) classification system, and (3) rates of use of the specific physician services (screening tests, diagnostic and imaging procedures) and major elective surgical procedures that are available in the *Dartmouth Atlas of Health Care*.¹⁹ We combined data from the fee-for-service (Part B) claims and the Medicare Provider Analysis and Review (MEDPAR) files.²⁰ Data for these analyses were all adjusted for the age, race, and sex composition of the population, and all spending measures were adjusted for inflation and differences in prices across states.²¹

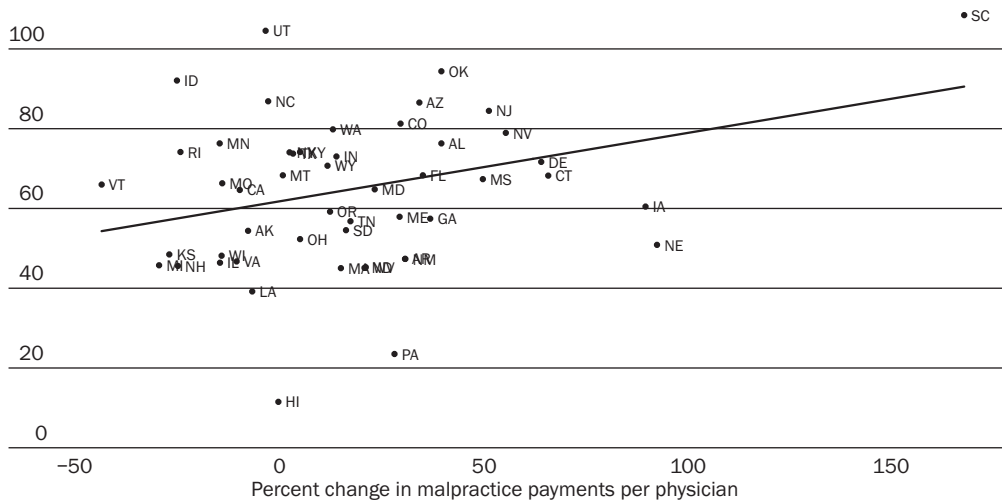
Study Results

Exhibit 1 reports summary statistics for our primary analysis. Between 1993 and 2001, total Medicare spending per beneficiary grew 35 percent and averaged \$6,500 in 2001 (all dollar figures are reported in 2001 dollars). Over the same time period, there was a 31 percent increase in spending on total physician services; the subcomponent of reimbursement for evaluation and management and imaging claims grew the most quickly. Between 1993 and 2001, mean malpractice payments per physician grew 12 percent, whereas malpractice premiums grew 8 percent. These national numbers mask considerable heterogeneity in growth rates across states.

■ **Malpractice payments and Medicare spending.** We first examined the simple association between malpractice payments per physician and two categories of Medicare spending of particular interest: Medicare spending on total physician services (Exhibit 2), and Medicare spending on the imaging subcomponent of physician services (Exhibit 3). These regressions suggest that increases in malpractice payments were associated with significant increases in Medicare spending on physicians in general and in spending on imaging in particular. These univariate regression results are quite consistent with the main multivariate regression results pre-

EXHIBIT 3
Longitudinal Association Between Growth In Malpractice Payments Per Physician And Medicare Spending On Imaging Services, 1993–2001

Percent growth in Medicare spending on imaging services



SOURCE: Authors' calculations.

NOTES: Univariate regression implies that a 10 percent increase in malpractice payments per physician is associated with a 1.73 percent (standard error = 0.74 percent) increase in spending. Regression line is population weighted. In 2001, average malpractice payments per physician were \$5,221.

main specification: a multivariate analysis controlling for both fixed state-specific factors and state characteristics that might change over time, such as population demographics and the economic climate. Exhibit 4 reports the regression-adjusted association between 10 percent growth within a state over time in our two liability measures and the growth of various Medicare spending components. These associations controlled both for any state-level characteristics of the malpractice environment or population and for the covariates noted above. Increases in payments per physician were statistically significant for spending on total physician services, the evaluation and management subcomponent, reimbursement for imaging services, and payments for minor surgical procedures. There was no statistically significant effect on the use of diagnostic procedures and major procedures. Thus, for example, a state with 10 percent higher growth in malpractice payments than its neighbor saw a little more than 1 percent higher growth in total spending on physician services, holding constant each state's idiosyncrasies as well as changes in the economic and demographic covariates. The second panel of Exhibit 4 reports results using premiums as an alternative measure of malpractice liability. These results are quite similar. Both measures of malpractice liability have a positive but statistically insignificant association with total Medicare spending. Specification tests using alternative models yielded strikingly similar

**EXHIBIT 4
Longitudinal Association Between Malpractice Liability And Medicare Spending,
1993–2001**

Measure of Medicare spending per beneficiary	Effect of 10 percent growth in malpractice payments per physician (N = 50 states)		Effect of 10 percent growth in malpractice premiums per physician (N = 50 states)	
	Percent increase	p value	Percent increase	p value
Total spending	0.6	0.18	1.0	0.17
Total physician (Medicare Part B) services	1.0	0.00	1.3	0.01
Evaluation and management	0.9	0.00	2.7	0.01
Diagnostic tests	0.0	0.95	0.2	0.77
Imaging	2.2	0.00	2.1	0.03
Minor procedures	1.0	0.01	0.9	0.03
Major procedures	-0.3	0.24	-0.1	0.90

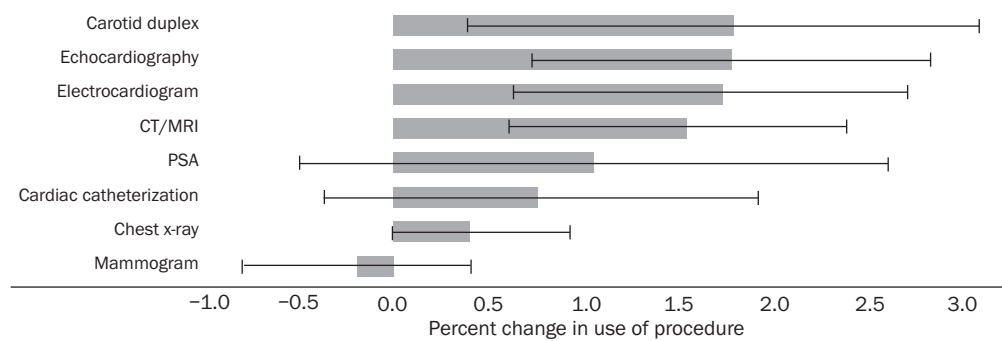
SOURCE: Authors' calculations, based on data below.

NOTES: Medicare spending data are from the *Dartmouth Atlas of Health Care* project, adjusted for age, race, and sex composition and inflation. Physician charges are classified using Berenson-Eggers Type of Service (BETOS) codes. Malpractice payments per physician were obtained from the National Practitioner Data Bank. Malpractice premiums per physician were obtained from the *Medical Liability Monitor*. "Percent increase" column reports the average percentage growth across all states. Dollar amounts are reported in 2001 dollars. Regressions are at the state level, weighted by state population, with heteroskedasticity-consistent standard errors. Covariates include per capita income, unemployment rate, percentage black, health maintenance organization (HMO) penetration, and percentage with high school degree. All variables were measured as the percentage growth within each state between 1993 and 2001.

results.²²

■ **Specific services and malpractice costs.** To identify the specific services that increase with higher malpractice costs, we studied the effect on procedure-specific utilization rates (Exhibits 5 and 6). A 10 percent increase in malpractice payments increases use of carotid duplex, echocardiography, electrocardiogram

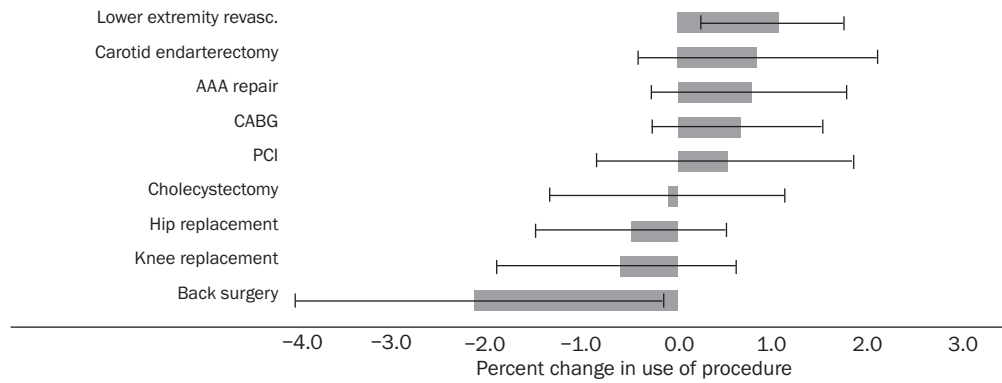
**EXHIBIT 5
Association Between A 10 Percent Increase In Malpractice Payments Per Physician
And The Use Of Diagnostic And Imaging Procedures**



SOURCE: Authors' calculations.

NOTES: Bars represent regression-adjusted coefficients from regression of log of procedure use on log of number of malpractice judgments or settlements per physician, with 95 percent confidence intervals indicated by the horizontal ruling lines. CT is computed tomography. MRI is magnetic resonance imaging. PSA is prostate-specific antigen.

EXHIBIT 6
Association Between A 10 Percent Increase In Malpractice Payments Per Physician
And The Use Of Major Surgical Procedures



SOURCE: Authors' calculations.

NOTES: Bars represent regression-adjusted coefficients from regression of log of procedure use on log of number of malpractice judgments or settlements per physician, with 95 percent confidence intervals indicated by the horizontal ruling lines. AAA is abdominal aortic aneurysm. CABG is coronary artery bypass graft. PCI is percutaneous coronary intervention.

(EKG), and computed tomography (CT)/magnetic resonance imaging (MRI) scanning by 1.5–1.8 percent ($p < 0.05$). Increases in the use of prostate-specific antigen (PSA) testing, cardiac catheterization, and chest x-rays were not significant at conventional levels. We did not find an increase in the use of mammograms (Exhibit 5). In contrast to the increased use of diagnostic and imaging procedures, we did not find that increases in malpractice costs were associated with increases in the use of major surgical procedures (Exhibit 6). The only exception to this finding is the use of lower extremity bypass—here, a 10 percent increase in malpractice payments was associated with a 1.0 percent increase ($p < 0.01$) in procedure use. In fact, there was a significant decrease in the use of back surgery, potentially the type of procedure that physicians might avoid when malpractice costs are higher. We found, however, only a statistically insignificant negative association between increased malpractice liability costs and overall spending on major procedures. This might be because the set of patients and conditions in which surgeons avoid surgery is relatively small.

There is always the possibility that confounders affected our analysis, so we performed several analyses to test the robustness of our findings. In particular, we were concerned about confounding variables that are positively correlated with premiums and payments and with diagnostic intensity but not with use of medical procedures. We studied the association between our measures of medical malpractice and the incidence of heart attacks, hip fractures, and mortality from cardiovascular disease and malignant neoplasms.²³ If there were a positive association between these variables, we would be concerned that an omitted variable such as population health could be driving both the increase in malpractice liability and the use of imaging services. The prevalence of neither heart attack nor hip fracture was affected by either of our measures of malpractice liability exposure.

“We did not find that higher malpractice liability costs were associated with reductions in total or disease-specific mortality.”

■ **Mortality and malpractice costs.** We also examined the association between mortality from various causes and our measures of malpractice costs.²⁴ Here, too, there was no significant association. This lack of correlation suggests not only that unmeasured changes in patient illness did not drive our results but also that the increased Medicare spending associated with rising malpractice costs did not measurably reduce mortality, although it certainly might have affected patient well-being in other ways.

Another possibility is that patients in some areas are becoming more “certainty oriented,” thereby explaining the use of diagnostic testing as well as an increase in litigation arising from allegations of failure to diagnose. To explore this hypothesis, we used data from a recent study and were unable to find evidence of geographic variation between census regions in patients’ preferences for routine cancer screening, free total-body CT scans, and the choice between receiving \$1,000 or a free body scan.²⁵ This result, although not definitive, is reassuring.

Finally, the NPDB specifies whether a malpractice payment was made for alleged malpractice in the areas of diagnosis, surgery, obstetrics, medication, equipment, anesthesia, or treatment. If the “certainty orientation” hypothesis were correct, we might expect an increase in payments associated with “failure to diagnose” and “delay in diagnosis” in states where malpractice liability increased. We found no evidence of such a relationship: The correlation between the percentage increase in malpractice payments per physician and the percentage increase in the number of diagnostic payments was 0.17 ($p < 0.24$). The correlation between the percentage increase in malpractice payments per physician and the percentage increase in the number of diagnostic payments in the narrower categories of “failure to diagnose” and “delay in diagnosis” was -0.03 ($p < 0.83$). Although neither of these tests irrefutably rejects the certainty-orientation hypothesis, they suggest that it was not a first-order source of bias.

Discussion

Our study used fairly recent data to estimate the association between increases in malpractice liability costs and changes in medical spending and practice patterns. We found that a 10 percent increase in average malpractice payments per physician within a state was associated with a 1.0 percent increase in Medicare payment for total physician services and a 2.2 percent increase in the imaging component of these services. We obtained similar results using malpractice premiums as an alternative measure of liability costs.

In addition to the increase in the use of imaging services, we saw a somewhat weaker increase in the use of other discretionary, generally low-risk services such

as physician visits and consultations, diagnostic tests, and minor procedures. A recent survey of physicians found that more than 93 percent ordered additional tests and performed additional diagnostic procedures in response to growing malpractice costs.²⁶ This survey also reported a substantial increase in the use of imaging technologies and a reduction in major surgeries among certain patient populations. Our results are consistent with these self-reports.

Our estimates shed some light on the magnitude of the relationship between malpractice liability and the use of medical services. States in the top quartile of malpractice payments per physician have 70 percent more payments per physician than states in the bottom quartile. Our estimates suggest that relative to states in the bottom quartile, all else equal, these states with high malpractice liability will have total Medicare spending that is 4.2 percent higher and spending on physicians that is 7.0 percent higher.

To put these estimates into perspective, consider the 60 percent increase in average malpractice premiums between 2000 and 2003. Our results suggest that this increase was associated with an increase Medicare spending of about \$16.5 billion total and \$7.1 billion on physician services (since Medicare outlays in 2003 were \$275 billion).²⁷

Although our analysis suggests an important association between malpractice costs and the use of imaging services, this link might have been missed in previous studies that focused on an earlier era, when the use of imaging procedures and outpatient services was less prevalent. Our estimates do not imply that any change in spending was necessarily “defensive medicine.” To the extent that additional malpractice costs mean greater precautionary testing with some medical value, any additional procedures might be protective of patient health or valued regardless of their therapeutic properties. We did not find that higher malpractice liability costs were associated with reductions in total or disease-specific mortality. This evidence is clearly not sufficient to rule out a potential benefit from malpractice liability-induced medical spending, but there is also some evidence from other studies that the increases in use associated with malpractice liability costs could actually lead to harm.²⁸

Our study is not without limitations. First, our sample was limited to the Medicare population; although this population accounts for a sizable share of overall health spending, our results might not generalize to other parts of the health care system. Second, although our longitudinal analysis was designed to account for all fixed unobservable confounders that operate at the state level and all national trends, unobserved confounders that vary within states over time might have affected our analysis. The specification tests we reported suggest that this was not the case, ruling out many of the most likely potential sources of bias, but outside of an experimental setting, it is difficult to prove causality conclusively.

.....
 This research was funded in part by the National Institute on Aging, NIA P01 AG19783-02. The opinions in this paper are those of the authors and should not be attributed to the NIA or any institution with which they are affiliated.

NOTES

1. M.M. Mello, D.M. Studdert, and T.A. Brennan, "The New Medical Malpractice Crisis," *New England Journal of Medicine* 348, no. 23 (2003): 2281–2284; and D.M. Studdert, M.M. Mello, and T.A. Brennan, "Medical Malpractice," *New England Journal of Medicine* 350, no. 3 (2004): 283–292.
2. T.A. Brennan et al., "Incidence of Adverse Events and Negligence in Hospitalized Patients: Results of the Harvard Medical Practice Study I," *New England Journal of Medicine* 324, no. 6 (1991): 370–376; and A.R. Localio et al., "Relation between Malpractice Claims and Adverse Events Due to Negligence: Results of the Harvard Medical Practice Study III," *New England Journal of Medicine* 325, no. 4 (1991): 245–251.
3. D.M. Studdert et al., "Defensive Medicine among High-Risk Specialist Physicians in a Volatile Malpractice Environment," *Journal of the American Medical Association* 293, no. 21 (2005): 2609–2617.
4. We used the term "malpractice liability costs" to refer to both malpractice insurance premiums and the size and number of malpractice judgments and settlements. As discussed below, the term is not intended to imply any specific causal relationship between these components of the malpractice liability environment and physician behavior or the value of services performed.
5. Studdert et al., "Defensive Medicine among High-Risk Specialist Physicians"; L.M. Baldwin et al., "Defensive Medicine and Obstetrics," *Journal of the American Medical Association* 274, no. 20 (1995): 1606–1610; A.R. Localio et al., "Relationship between Malpractice Claims and Cesarean Delivery," *Journal of the American Medical Association* 269, no. 3 (1993): 366–373; L. Dubay, R. Kaestner, and T. Waidmann, "The Impact of Malpractice Fears on Cesarean Section Rates," *Journal of Health Economics* 18, no. 4 (1999): 491–522; U.S. Congress Office of Technology Assessment, *Defensive Medicine and Medical Malpractice*, Report no. OTA-H-602 (Washington: OTA, 1994); and D.P. Kessler and M.B. McClellan, "Do Doctors Practice Defensive Medicine?" *Quarterly Journal of Economics* 111, no. 2 (1996): 353–390.
6. Medicare Payment Advisory Commission, *Healthcare Spending and the Medicare Program* (Washington: MedPAC, 2005); and J.K. Iglehart, "The New Era of Medical Imaging—Progress and Pitfalls," *New England Journal of Medicine* 354, no. 26 (2006): 2822–2828.
7. D. Verrilli and H.G. Welch, "The Impact of Diagnostic Testing on Therapeutic Interventions," *Journal of the American Medical Association* 275, no. 15 (1996): 1189–1191.
8. D.P. Kessler, W.M. Sage, and D.J. Becker, "Impact of Malpractice Reforms on the Supply of Physician Services," *Journal of the American Medical Association* 293, no. 21 (2005): 2618–2625; and K. Baicker and A. Chandra, "The Effect of Malpractice Liability on the Delivery of Health Care," in *Frontiers of Health Policy Research*, ed. D. Cutler and A.M. Garber (Cambridge, Mass.: MIT Press, 2005), 16–18.
9. Kessler and McClellan, "Do Doctors Practice Defensive Medicine?"
10. We weighted each state according to its population in the 2000 census (so that results can be interpreted as applying to the average person).
11. Kessler and McClellan, "Do Doctors Practice Defensive Medicine?"; and D.P. Kessler and M.B. McClellan, "How Liability Law Affects Medical Productivity," *Journal of Health Economics* 21, no. 6 (2002): 931–955.
12. Z. Griliches and J.A. Hausman, "Errors in Variables in Panel Data: A Note with an Example," *Journal of Econometrics* 31, no. 1 (1985): 93–118.
13. National Center for Health Workforce Analysis, Area Resource File (Rockville, Md.: Health Resources and Services Administration, 2003). Sensitivity to these choices, discussion of other potentially omitted factors, and estimates with additional controls are included in an online appendix, available at <http://content.healthaffairs.org/cgi/content/full/26/3/841/DC1>.
14. Kessler and McClellan, "How Liability Law Affects Medical Productivity."
15. A. Chandra, S. Nundy, and S.A. Seabury, "The Growth of Physician Malpractice Payments: Evidence from the National Practitioner Data Bank," *Health Affairs* 24 (2005): w240–w249 (published online 31 May 2005; 10.1377/hlthaff.w5.240).
16. B. Black et al., "Stability, Not Crisis: Medical Malpractice Claim Outcomes in Texas, 1988–2002," *Journal of Empirical Legal Studies* 2, no. 2 (2005): 207.
17. Baicker and Chandra, "The Effect of Malpractice Liability"; and Chandra et al., "The Growth of Physician Malpractice Payments."

18. Federal Reserve Bank of St. Louis, "Gross Domestic Product: Implicit Price Deflator," 2004, <http://research.stlouisfed.org/fred2/data/GDPDEF.txt> (accessed 14 February 2007).
19. Berenson-Eggers Type of Service (BETOS) Codes, 2005, http://www.cms.hhs.gov/HCPSCReleaseCodeSets/20_BETOS.asp (accessed 12 March 2007); and J. Wennberg and M. Cooper, *The Dartmouth Atlas of Health Care* (Chicago: American Hospital Association Press, 1999). This study, and its underlying protocol guaranteeing the confidentiality of the Medicare claims data, was approved by the Institutional Review Board (IRB) at Dartmouth College.
20. A 5 percent sample of Medicare fee-for-service physician (Part B) claims was used to calculate age-, race-, and sex-adjusted rates of spending on total physician services and for each of the major BETOS categories. Total Medicare spending per beneficiary was also ascertained from the same 5 percent sample, using records from the Continuous Medical History Sample File. Rates of major elective inpatient surgical procedures were based upon a 100 percent sample drawn from the Medicare Provider Analysis and Review (MEDPAR) file, and rates of specific physician services were calculated from a 20 percent sample of Part B physician claims in later years and a 5 percent sample in earlier years. The population denominator for all rates was the midyear population of fee-for-service Medicare beneficiaries, age sixty-five and older, who were eligible for both Parts A and B.
21. We used a state-level cost-of-living adjustment to adjust all premium, payment, and spending dollar values for state-level variation in prices, although as shown in the appendix, this does not affect subsequent regression results. See Note 13.
22. In the appendix exhibits we report a number of specification tests, including results from models using two alternative sets of weights (state population from the 1990 census and the number of physicians in each state) as well as including other covariates. See Note 13.
23. See Appendix Exhibit 3; *ibid.*
24. *Ibid.*
25. L.M. Schwartz et al., "Enthusiasm for Cancer Screening in the United States," *Journal of the American Medical Association* 291, no. 1 (2004): 71–78. Geographic identifiers in this study were limited to the four major census regions. We performed a chi-square test to examine if there were geographic differences in preferences for screening.
26. Studdert et al., "Defensive Medicine among High-Risk Specialist Physicians."
27. Congressional Budget Office, *The Budget and Economic Outlook: Fiscal Years 2005 to 2014*, January 2004, <http://www.cbo.gov/showdoc.cfm?index=4985&sequence=0&from=0#anchor> (accessed 14 February 2007). We focused on the responsiveness of health care spending to malpractice liability in the Medicare population. There is evidence that elderly beneficiaries are much less likely than others to litigate, which suggests that our analysis might understate the response in the general population. However, most beneficiaries are enrolled in fee-for-service, where, unlike capitated plans, there are few restrictions on a physician's ability to order additional tests—a possibility that suggests that results from Medicare might be larger than the economywide responsiveness of physicians to malpractice costs. If these effects roughly offset each other, extrapolating these estimates to the general population would suggest that the 60 percent increase in malpractice premiums between 2000 and 2003 would be associated with a 6 percent, or \$95 billion, increase in national health spending. Given that our data drew only from the Medicare population, however, the true effect on national health spending might be quite different.
28. E.S. Fisher and H.G. Welch, "Avoiding the Unintended Consequences of Growth in Medical Care: How Might More Be Worse?" *Journal of the American Medical Association* 281, no. 5 (1999): 446–453.