

## OUTCOMES ASSESSMENT AND HEALTH CARE REFORM

Jonathan Skinner<sup>1,2,3</sup>

Amitabh Chandra<sup>1,3,4</sup>

Elliott Fisher<sup>1,5,6</sup>

Draft v12: February 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.

<sup>1</sup> Community and Family Medicine, Dartmouth Medical School, Hanover NH; <sup>2</sup>Department of Economics, Dartmouth College, <sup>3</sup> National Bureau of Economic Research, Cambridge, MA; <sup>4</sup> Kennedy School of Government, Harvard University, Cambridge MA; <sup>5</sup> Department of Medicine, Dartmouth Medical School, <sup>6</sup> 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.<sup>1</sup> 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

---

<sup>1</sup> 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  $\beta$  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.<sup>2</sup> 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.<sup>3</sup> 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

<sup>2</sup> 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.

<sup>3</sup> 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

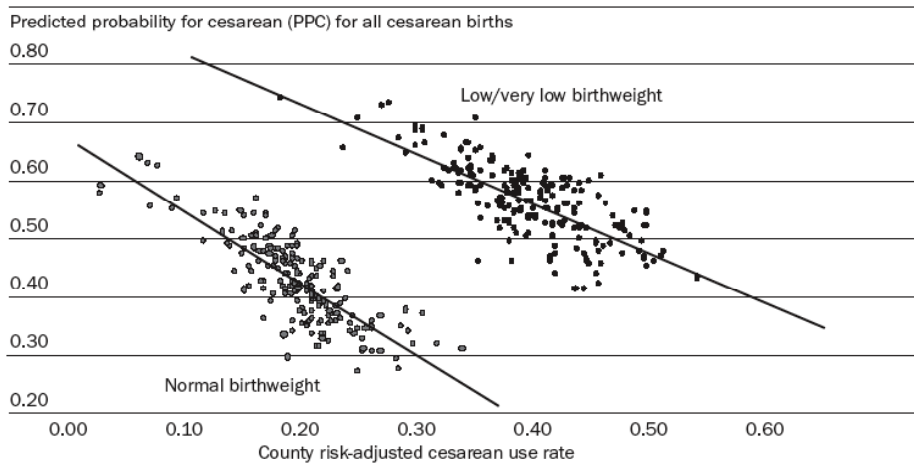
Whether or not a procedure is therapeutic fundamentally depends on a patient’s clinical status. Even for patients within a clinical trial, there are substantial differences in the relative benefits across risk strata, with some patients receiving no benefit (or even harm), while a smaller group achieves 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).

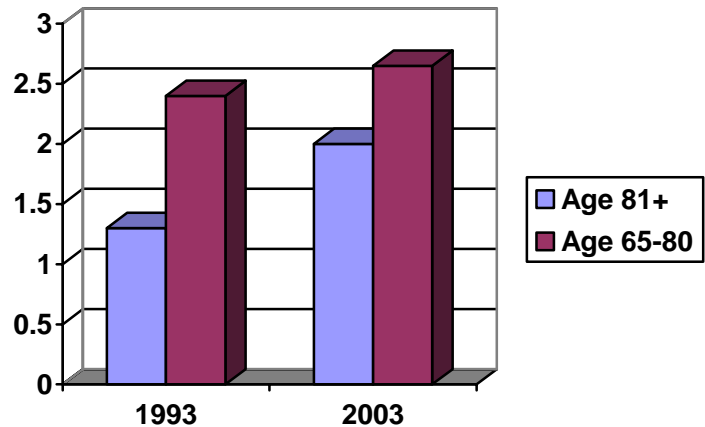
**Figure 1**

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.<sup>4</sup>

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

---

<sup>4</sup> 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.<sup>5</sup>

### *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

---

<sup>5</sup> 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.<sup>6</sup> 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  $\beta$  blockers and aspirin within 24 hours of the AMI.

---

<sup>6</sup> 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  $\beta$  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-

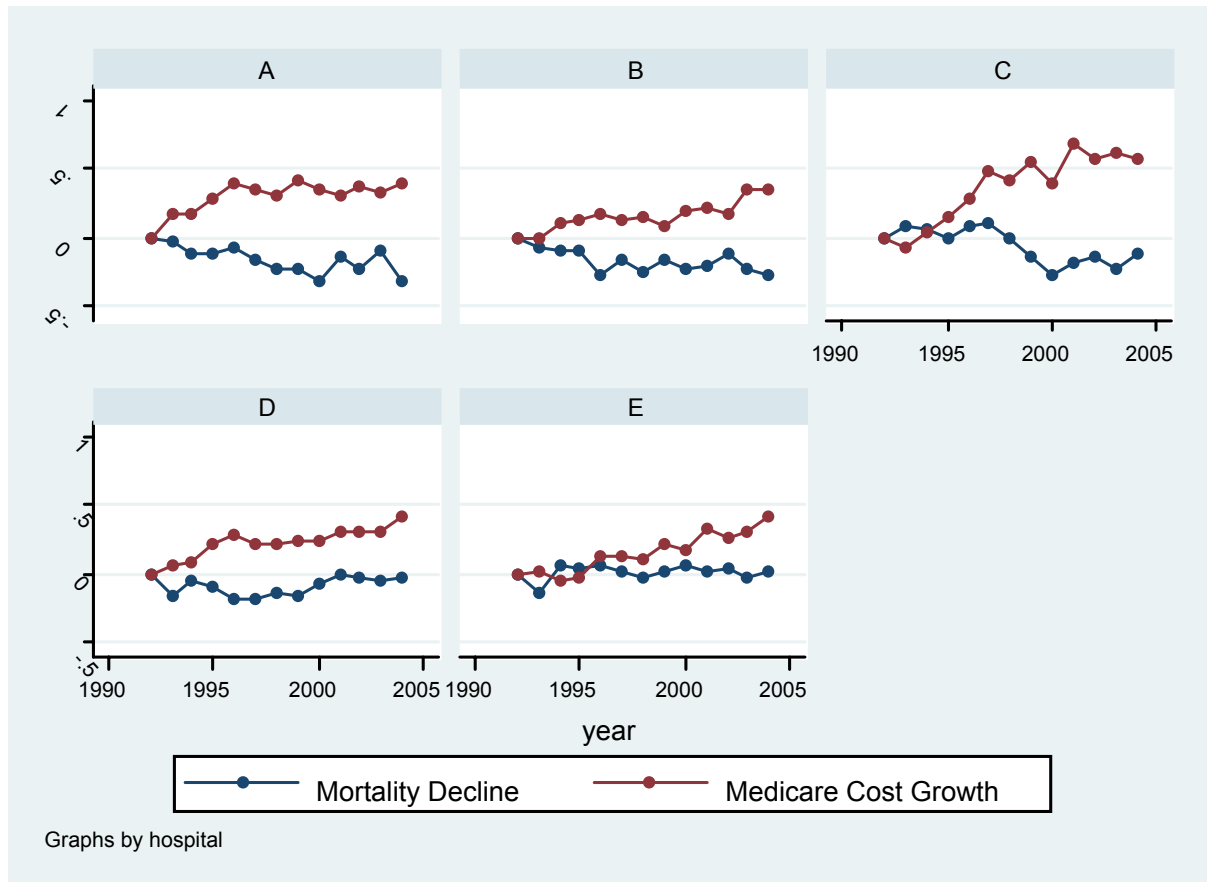
	Average*	Hosp A	Hosp B	Hosp C	Hosp D	Hosp E
Adj. 1 Year Mortality, 1992	0.346	0.366	0.415	0.326	0.361	0.291
Adj. 1 Year Mortality, 2004	0.297	0.250	0.305	0.289	0.356	0.294
Mortality Diff.	-0.049	-0.116	-0.110	-0.037	-0.005	0.003
Adj. 1 Year Expenditures, 1992	19,991	14,785	16,492	22,961	18,799	15,425
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,  $\beta$  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  $\beta$  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  $\beta$  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.<sup>7</sup> 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.

---

<sup>7</sup> 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 database<sup>8</sup>. 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.

---

<sup>8</sup> 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.<sup>9</sup>

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.<sup>10</sup> 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

---

<sup>9</sup> 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.

<sup>10</sup> 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,<sup>11</sup> 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.

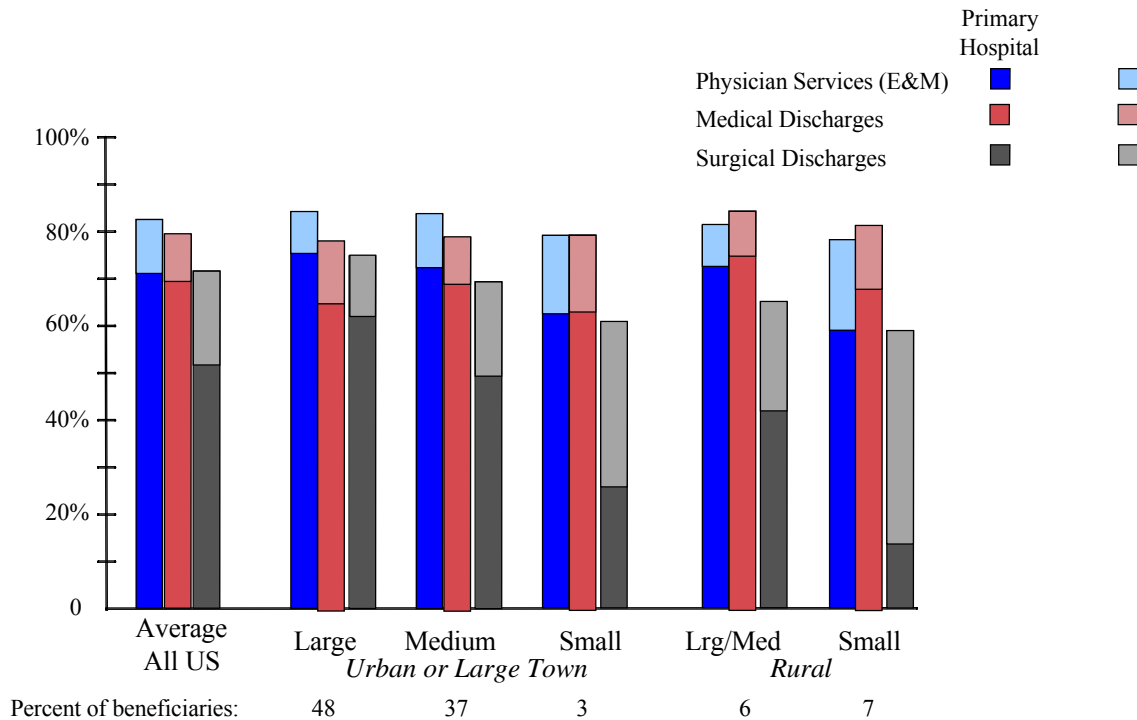
#### ***4.3. Translating incentives into practice: The Accountable Care Organization***

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

---

<sup>11</sup> 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% pf MDs to a single acute care hospital in the US.<sup>12</sup> 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

<sup>12</sup> 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.<sup>13</sup> 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

---

<sup>13</sup> 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. Gøtzsche, 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.