RAND

Reimbursement Under Uncertainty

What to Do If One Cannot Identify an Efficient Hospital

Joseph P. Newhouse

RAND/UCLA/Harvard Center for Health Care Financing Policy Research

The research described in this report was sponsored by the Health Care Financing Administration, U.S. Department of Health and Human Services, under Cooperative Agreement No. 99-C-98489-9-08.

Library of Congress Cataloging in Publication Data

Newhouse, Joseph P.
Reimbursement under uncertainty : what to do if one cannot identify an efficient hospital / Joseph P. Newhouse.
p. cm.
"Supported by the Health Care Financing Administration."
"MR-309-HCFA."
Includes bibliographical references.
ISBN 0-8330-1493-5
1. Hospital care—United States—Cost effectiveness—Econometric models. 2. Hospitals—United States—Cost of operation—Econometric models. 3. Insurance, Hospitalization—United States—Econometric models. I. United States. Health Care Financing Administration. II. Title.
RA981.A2N484 1994
338.4'336211'0973—dc20 93-46638
CIP

RAND is a nonprofit institution that seeks to improve public policy through research and analysis. RAND's publications do not necessarily reflect the opinions or policies of its research sponsors.

Published 1994 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
To obtain information about RAND studies or to order documents,
call Distribution Services, (310) 451-7002

RAND

Reimbursement Under Uncertainty

What to Do If One Cannot Identify an Efficient Hospital

Joseph P. Newhouse

Supported by the
Health Care Financing Administration,
U.S. Department of Health and Human Services

RAND/UCLA/Harvard Center for
Health Care Financing Policy Research

Preface

This report examines methods for identifying efficient hospitals for the Health Care Financing Administration. The research reported here was performed within the RAND/UCLA/Harvard Center for Health Care Financing Policy Research.

Contents

Preface iii

Summary vii

Acknowledgments xi

1. INTRODUCTION 1

2. DIFFICULTIES IN IDENTIFYING AN EFFICIENT HOSPITAL 2

3. EFFORTS TO ADJUST FOR OUTPUT DIFFERENCES 5
   How Much Information Is There? 10
   Is the Hospital Too Aggregate a Unit? 13

4. CONCLUSIONS 15
   A Multiple Payer System 19
   Recommendations 20

Appendix
A. THE BASIS OF REIMBURSING HEALTH CARE PROVIDERS 25

References 57

Summary

The purpose of this project is to comment on methods for identifying an efficient hospital, because the Boren Amendment requires the Medicaid program to reimburse hospitals at a rate that is "reasonable and adequate to meet the costs that must be incurred by efficiently and economically operated facilities." Satisfying this mandate would thus appear to require the identification of efficient facilities and calculating their costs.

A direct approach to this task is to estimate frontier cost or production functions; i.e., determine hospitals that, for a given output, seem to be minimizing cost; as a byproduct one obtains an estimate of the degree to which other hospitals lie inside the frontier, that is, are inefficient.

The rub comes in the measurement. Although the more serious issues concern output measurement, there are rather mundane issues even with respect to measuring costs and inputs; e.g., capital, physician, and contract nurse inputs are often omitted.

The existing literature usually assumes that the output of hospitals is simply patient days or stays or both, neglecting both case mix variation and the quality of the services delivered. Sometimes the Medicare Case Mix Index is included to adjust for case mix variation; although presumably better than not adjusting, the Medicare Case Mix Index is imperfectly correlated with case mix variation among non-Medicare cases and is an inexact adjuster even for the Medicare cases because of heterogeneity within diagnosis-related groups (DRGs). More generally, distinguishing higher costs due to inefficiency from those due to different case mix, or a different quality output, is a central and serious problem.

Additionally, given the amount of available data, strong assumptions must be made to estimate these functions. A standard econometric technique for making minimal assumptions about functional form is to use flexible functional forms. The price of minimal assumptions on functional form, however, is the estimation of many parameters; if the sum of inputs and outputs is only 100—markedly fewer than even the number of DRGs—a flexible functional form would require estimating over 5,000 parameters, but this is approximately the number of hospitals in the country. Moreover, in its simplest form this technique requires positive levels of output and input in each hospital, but some hospitals do not produce some outputs or product lines. Empirically a standard modification, the

generalized translog function yields unstable results as specific input or output levels approach zero.

Accommodating these problems has led the literature to use "hybrid" functional forms and to constrain severely the number of inputs and outputs considered; that is, to impose a great deal of structure and aggregation. Typically, those estimating hospital cost or production functions in the literature employ fewer than 100 parameters; nonetheless, they typically explain over 90 percent of the variance in cost across hospitals. This suggests that there is insufficient independent variation in the data to estimate the 1,000 or more parameters that would be required to estimate flexible functional forms even if the sum of inputs and outputs were constrained to be a little over 30.

Although the lack of independent variation could be taken to mean that estimation of additional parameters is not necessary because various inputs or outputs move together, it is a strong assumption that this covariance reflects a technological relationship; i.e., that the greater within-category heterogeneity, which would result from aggregating outputs or inputs, would not be exploited in an actual reimbursement scheme if it were profitable to do so. Consider, for example, the potential distortions in the Prospective Payment System (PPS) that might arise if the number of DRGs were reduced from 475 to 100. Even though the variation explained by the DRGs in an existing data set may not fall much from the reduction to 100 DRGs, within-DRG variance could substantially increase in a number of instances, thereby enhancing opportunities for patient selection.

Furthermore, both the legislative language to identify efficient hospitals and the existing empirical literature assume that a hospital is equally efficient or inefficient in all its endeavors. A moment's reflection will indicate this is not likely to be the case; the physicians in one department may be more skilled (or obtain better results) than in another; the pharmacy may be badly managed, whereas the laboratory is well managed. The spirit of the Boren Amendment suggests that costs of particular units be estimated; the hospital might be too aggregate an enterprise.

Analysis at the department or case level, however, poses severe problems. Allocation of joint cost becomes an issue, as do economies of scope. Moreover, reimbursement of hospitals at a level that reflected least-observed cost for each department might well drive all existing firms out of business.

Because of the foregoing problems I conclude that the prospects for identifying efficient hospitals through direct estimation of cost and production functions are not good. Perhaps such functions can be used as a screen to identify outliers,

which would receive closer attention; it is not clear, however, that even closer attention can distinguish inefficiency from "justified" higher costs, where justified is in quotes because it is not clear what is justified.

Does the economics literature then have anything relevant to say about rate setting? The industrial organization literature on government procurement, as well as the health economics literature, has considered the problem of rate setting when the regulator is uncertain about what costs would be if the firm made the optimal effort to be efficient. In general, this literature concludes that reimbursement should be partly a function of the firm's observed cost and not purely prospective, as is now the case in Medicare. Under some strong assumptions, the health economics literature suggests that 15 to 30 percent of hospital costs should be reimbursed, with the remainder of the reimbursement being set independently of the hospital's costs (but in all likelihood a function of some group of other hospitals' costs). I review this literature in the appendix. Because the literature is inconclusive on the weight to be attached to hospital costs, I suggest experimentation with "partial capitation."

Even if one agrees that reimbursement should be partly a function of a firm's observed costs, that is not sufficient to determine a level of payment. Although conceptually messy, I suspect the preferable alternative to basing rates on empirically estimated frontier functions is to engage in an incremental, empirical approach, much like what the United States and other countries have engaged in historically, where the regulator makes incremental changes and observes the consequences. If the consequences move in the desired direction, the regulator makes a further incremental change. This implies the need for timely and reasonably complete information on output; i.e., for investment in monitoring systems. Use of this strategy assumes that hospitals cannot collude or act strategically against the regulator. At a national level, such an assumption seems reasonable; if the regulator is in a local market, such an assumption is less tenable.

A prominent but distinct issue in the current debate is whether Medicaid underpays. Those arguing that it does (or might if payment falls further) implicitly assume that it is desirable for patients to be treated in one type of hospital with one quality level rather than being segregated in different facilities with different quality levels. Assume for the sake of argument that nonsegregation is desirable (indeed, having different facilities at different quality levels could conceivably increase social cost if fixed costs for a facility are high enough).

Then in a simple model with two possible levels of quality (Medicaid and private), one can show that if the Medicaid payment remains above marginal cost at the higher quality level, patients will not be segregated. If not, and if hospitals do not cross subsidize, one might find segregation. Because it is easier to monitor the concentration of Medicaid (and Medicare) beneficiaries across hospitals, controlling for the geographic distribution of beneficiaries, than it is to estimate marginal cost, I suggest such monitoring, assuming segregation is undesirable.

Acknowledgments

I would like to thank Tony Bower, Susan Ettner, Steve Garber, Jonathan Skinner, and Richard Zeckhauser for helpful comments on a preliminary draft of the report. David Cutler helped stimulate my thinking with some early discussions of the material in the appendix. I am grateful to all of them; any remaining deficiencies are my responsibility, especially since I sometimes chose not to follow their advice!

1. Introduction

The objective specified by the Health Care Financing Administration for this project was to develop a method for identifying an efficient hospital. The need to do so follows from legislation, commonly referred to as the Boren Amendment, that requires the Medicaid program to reimburse hospitals at a rate that is "reasonable and adequate to meet the costs that must be incurred by efficiently and economically operated facilities." Satisfying this mandate would thus appear to require the identification of efficient facilities and calculating their costs.

Unfortunately for the objective, I will argue that the data and measurement methodology to identify efficient hospitals are not sufficiently developed to use for rate-setting purposes. Moreover, the problems seem so intractable that there appears little likelihood the situation will markedly change anytime soon. Thus, the relevant issue is how to reimburse hospitals given that this is the case. The same problem arises in other regulated industries, and the report draws on the industrial organization literature that addresses this problem. I begin, however, by considering the problems of trying to estimate directly the costs of an efficient hospital and the literature on that subject.

2. Difficulties in Identifying an Efficient Hospital

A fundamental problem in measuring efficiency is measuring the hospital's outputs to adjust for output differences among hospitals. Typically the empirical literature in health economics has taken the output of the hospital to be days or stays in some time period, say a year. But annual hospital days per person are lower now than they were in 1960, which, if days were truly the output, suggests that the more than fourfold increase in real spending per person since 1960 has been pure inefficiency.[1] No one could seriously put such a proposition forward, given the change in the nature of treatment for many diseases. At least some, and I have argued elsewhere a good bit, of the increase must be attributable to these changes (Newhouse 1992, 1993); that is, to changes in the nature of the product produced by the hospital. Although variation in technology over time is not relevant for the cross-sectional estimation that these studies typically engage in, it is clear that there is substantial variation in the cross section too; indeed, the term tertiary-care (and even quaternary-care) hospital is used to distinguish some hospitals from others.

Although the distinction is analytically not important, the change in the product has occurred—and occurred differentially across hospitals—for both the technical methods of care and for amenities; some examples may help fix ideas. The treatment of acute myocardial infarction has changed with the advent of coronary care units, catheterization, revascularization, and thrombolytic agents. The treatment of diseases of joints has changed with the development of arthroscopic surgery and artificial hips and knees; the treatment of premature infants has changed with the development of neonatal intensive care units; and the treatment of genetic disease has begun to change with products such as Ceredase for Gaucher's Disease and promises to change more in the next decade with the development of other genetically engineered biotechnology products. Indeed, a biomedical research enterprise has been devoted to finding treatments

[1]The factor of four figure in the text is derived using a measure of general inflation, the gross domestic product (GDP) deflator, rather than a hospital-specific price index. To the degree that hospital productivity lags, one would expect a hospital-specific price index to rise faster than the GDP deflator; thus, a portion of the fourfold expenditure increase could reflect lagging productivity rather than inefficiency (Baumol, 1992). I argue here, as I have argued elsewhere, that because of new products we do not have reliable measures of price change (Newhouse 1992); thus, I think it is impossible to know how much of the expenditure increase reflects lagging productivity.

for previously untreatable diseases or improving treatments of existing diseases; these treatments may or may not be available at all hospitals.

An example of changed amenities, now rather dated, is the shift from wards to semiprivate rooms, as well as the general newness of many hospital facilities. Indeed, a common complaint about hospital costs is overinvestment in structures, but undoubtedly for some patients and their families a newer building has some value (i.e., the associated spending does not represent pure inefficiency).

Other changes in treatment are less easy to classify as technical or amenities. For example, even if a device allowing the patient to self-regulate pain medication has little or no effect on the postdischarge outcome, a patient in pain waiting for a nurse may not agree that the device is simply an amenity. Nurse staffing levels in hospitals have generally increased; on the whole, it is reasonable to think that this has improved response times, which may improve both clinical outcome and patient satisfaction.

The nurse staffing example illustrates a key problem with the literature's use of the patient day or stay as the measure of output in cross-sectional data. Consider two hospitals that are otherwise alike with respect to inputs and cases, and in particular have the same number of stays and days, but one has 10 percent more nurses than another. One would be rash to conclude based on only this bit of information that the hospital with more nurses is 10 percent less efficient. As we shall see, however, most all existing methods for measuring efficient hospitals in effect do this or something close to this.

An analogy will illustrate the problem of doing so. Consider the additional flight attendants per passenger in the first-class section of an airplane. Undoubtedly these attendants provide the average first-class passenger with more attention than the average coach-fare passenger receives. If this attention were not valued, one presumes airlines would have the same flight attendants serving passengers in both first class and coach—with roughly the same amount of attention to each—rather than separate sets of attendants for each section. Thus, one should not conclude that the additional costs of flight attendants per passenger mile in first class represent inefficiency.

It seems likely a similar phenomenon operates in hospitals. Patients undoubtedly value faster nurse response times and may value additional time spent conversing with nurses. There is conceptually an optimal number of nurses (given by the usual first-order conditions), but concluding that the minimal observed number of nurses per day or stay is efficient—and that

hospitals with additional nurses per patient day or stay are simply inefficient—is almost certainly wrong.

One could argue that the hospital market is distorted by insurance, such that the subsidy from insurance induces demand for too many nurses per patient and too many resources generally. Thus, a lower level of nurses may be preferred by individuals in their role as premium payers or taxpayers rather than in their role as patients. However, unlike the quantity margin (e.g., admissions) we do not know much about moral hazard on the quality dimension (Manning et al., 1987), and what evidence there is suggests there may be much less moral hazard on this dimension. Marquis (1985), for example, finds that the amount of cost sharing did not much affect patient choice of physician (i.e., costliness of the physician chosen or specialty). Thus, the presumption that hospitals using fewer resources per patient day or stay are closer to a social optimum because of less moral hazard may not be correct. Furthermore, patients' preferences are surely heterogeneous, and some patients could be willing to pay for more than the minimum observed staffing level; hence, heterogeneity in staffing levels may be—and probably is—efficient rather than inefficient.[2]

---

[2]I assume, as is the case for Medicare and Medicaid, that balance billing by hospitals is prohibited, so a market solution to heterogeneous tastes is not available.

3. Efforts to Adjust for Output Differences

As the nurse staffing example illustrates, a central problem in the measurement of hospital efficiency is the measurement of output. Although one can discuss efficiency only in the context of producing a given output, the empirical measures of output tend to be simplistic. As already noted, most examples in the literature purporting to demonstrate the (in)efficiency of hospitals simply use the patient day or the patient stay as a measure of output.

These studies either estimate production functions or their dual, cost functions. An early example was Martin Feldstein's doctoral dissertation and 1968 book, which estimated a production function for hospitals in the British National Health Service using cases treated and length of stay as outputs (Feldstein 1968). Although Feldstein's analysis was at the time a major methodological advance, the measurement of output has not made much progress since.

Wagstaff (1989) gives a survey of British studies on this point; a now somewhat dated survey of American cost function studies is contained in Cowing, Holtmann, and Powers, (1983). More recent American studies in this genre are Granneman, Brown, and Pauly (1986), Jensen and Morrisey (1986a, 1986b), Vitaliano (1987), and Vita (1990).

I focus on two studies that seek to measure inefficiency in hospital production because they are good examples of two different methods and serve to illustrate the problems with each method. Grosskopf and Valdmanis (1987) use the technique of data envelopment analysis (DEA), which has been in the general economics literature since 1957. Stripped to its essentials, this technique finds hospitals that use the least amount of inputs for a given output and deems these efficient. In the Grosskopf-Valdmanis study output is measured as days, surgeries, or outpatient visits; thus, a hospital with more nurses per patient day than another (other things being equal) would register as proportionately more inefficient.

A second, more recent study attempts to go beyond days or stays as an output measure (Zuckerman, Hadley, and Iezzoni, 1993). Additionally, the authors of this paper seek to measure the amount of inefficiency in the production of hospital services using stochastic frontier estimation. This method represents a potential methodological advance over the deterministic estimation of DEA because it allows for random error in the observation of the dependent variable

(cost in a cost function, output in a production function), and random error surely exists in any real-world data set.[1] A central issue, however, is how completely Zuckerman et al. succeed in measuring output.

In fact, Zuckerman et al.'s measures are more detailed than those previously used in this literature. Moreover, and importantly, they have included measures of quality: very high or very low 30-day mortality rates; and measures that suggest problems with quality of care (based on actual versus expected rates of occurrence of events such as wound infections and medications incidents, insofar as those are recorded in claims data). Zuckerman et al., conclude that around 14 percent of hospital spending is inefficient, in the sense that 14 percent fewer inputs could produce the same output mix.[2]

Because Zuckerman et al. is in my view the best study of this genre to date, I use it to illustrate just how difficult the problem is of identifying inefficiency:

1. Most fundamentally, all aspects of output are still not included. For example, the Zuckerman et al. analysis would likely label the addition of a nurse who responded more quickly to patients and was less harried in dealing with them as inefficiency. The nurse would be considered useful only if her hiring had an effect on mortality rates or reduced the recorded incidents with quality of care (or permitted the treatment of patients in diagnosis-related groups [DRGs] with higher case weights). Thus, one cannot rule out from the results the possibility that there is no inefficiency at all in hospital production, implausible though this is.
2. For those aspects of quality that are included, one must determine for what level of quality one is willing to pay; i.e., at what quality level would one find an efficient hospital?[3] For example, in the Zuckerman et al. study, cost is conditional upon the level of recorded incidents involving quality of care. Thus, the estimates in that study at best define a frontier showing, for example, the minimum number of incidents associated with a given cost; the regulator is then left to pick a cost on that frontier. What is the proper

---

[1]Stochastic frontier cost estimation, however, does require parametric assumptions about the nature of the random error. These assumptions usually do not come from theory but are rather made for mathematical tractability. Zuckerman et al. argue that their results are not sensitive to alternative parametric assumptions; their argument, however, is only partially correct; see the text below.

[2]The 14 percent value appears to be in logarithmic space; if the figure were in raw dollar space, it is not clear whether it would be higher or lower. The standard deviation of this value is around 6 percent. Estimated inefficiency at the hospital level is sensitive to observed relative to expected salaries and is moderately sensitive to staff per adjusted admission and occupancy rates.

[3]The issue of the optimal quality level is inherent in the choice problem and is not a measurement issue; I include it here, however, because the Boren Amendment ignores the choice issue.

number of untoward incidents? Some may be tempted to respond zero, but zero is unlikely to be optimal if, as with pollution, eliminating the last few incidents is exceedingly expensive.[4]

In fact, the problem is worse to the degree that incidents are not always observed and their frequency is affected by staffing; in such a case, the true trade-off of cost and quality will differ from the estimated trade-off.

3. The technique cannot detect inefficiency among the best performing (frontier) hospitals; the assumption is that the best performing are efficient or at least one cannot for practical purposes do better. Although this assumption may seem innocuous, the United States has high staff/bed ratios relative to most other developed countries [e.g., personnel per bed in the United States were 3.35 in 1990, versus 2.40 in Canada in 1988 (latest year) and 1.31 in Germany in 1989 (latest year); the only developed country with a higher value is Australia with 3.91 in 1989 (Schieber, Poullier, and Greenwald, 1992)], raising the possibility that all U.S. hospitals may be off a frontier.[5] Put another way, even the best performing American hospitals might be able to improve their performance.

4. If efficiency comes from greater scale, the literature does not generally consider any higher travel costs that may be imposed on patients or their families from utilizing larger but fewer facilities, but such costs should be considered in arriving at a social optimum.

5. Using the stochastic frontier estimation method, any given measure of inefficiency is necessarily dependent upon the parametric assumptions made about the unobserved errors in the dependent variable. (With DEA the measure is necessarily dependent upon the assumption of no measurement error in the dependent variable and the assumption that any conditions that gave rise to low costs are replicable.)[6] Although Zuckerman et al. argue that their measure of inefficiency is insensitive to the particular parametric variations they tried, they varied only the parameterization of the

---

[4]A good estimate of the cost of eliminating the last few incidents is probably not available. Although Zuckerman et al. do not tell us, it is likely very few or no hospitals of any size have zero incidents in a year, so that any estimate of the cost of reducing to zero is likely to require extrapolating beyond the range of the data.

[5]An alternative interpretation is that because U.S. stays are shorter, the average patient is sicker, and more personnel is required. Haber et al. (1992), however, show that factor shares of clinical and nonclinical personnel in the U.S. and Canada are about the same; it is not clear why sicker patients would necessitate more food service helpers and cleaners.

[6]Replicability is more of an issue with DEA than with standard methods because in standard methods idiosyncratic firm-specific effects do not affect the regression coefficients provided the usual assumption of independence between the error and the covariates is satisfied. In DEA, however, idiosyncratic elements in one direction will define the frontier.

inefficiency factor, not the underlying error structure in the cost function.[7] In their method the underlying error in the cost function is assumed to be normal (actually, normal in the logarithm, since they use the log of cost as a dependent variable). The identification of inefficiency then results from the skewness that the inefficiency component (which they assume to be half normal) imparts to the overall error term. If, however, the underlying error in the cost function is not normal, then the inefficiency cannot be identified from the skewness of the overall error term. It is hard to develop intuition about the distribution of the error in the underlying cost function, but the theoretical rationale for a normal error, presumably an appeal to the central limit theorem, appears weak. The problem is analogous to identifying sample selection bias from parametric assumptions about the error (e.g., normality), a situation that arises when one does not have an omitted variable that influences sample selection but does not influence the underlying relationship one is trying to estimate.

These untestable assumptions on the error term carry a large burden because measurement error will surely be considerable.[8] For example, to allow for the possible endogeneity of factor prices, Zuckerman et al. use an instrumental variable for factor prices, predicted actual salary per person as a function of the size of the Metropolitan Statistical Area (MSA), census division, and hospital control. This is analogous to using the mean salary for a cell for each hospital in the cell. Thus, this method throws the difference between the observed salary level at the hospital and the mean salary level for the cell into the error term of the cost function, but this difference need not be normally distributed. To the degree it is not, Zuckerman et al.'s finding that the difference between observed and expected salaries predicts inefficiency at the hospital level is suspect.

I note in passing that the factor price measurement problem extends on to a regulator's problem in implementing any directive to price at an efficient level because to set a budget the regulator must know hospital-specific factor prices. Suppose hospitals are price takers, in principle an easier problem for the regulator because he or she does not then need to worry about possible

---

[7]In this method the observed error is assumed to be made up of two additive components. One component is called in the text the underlying error in the cost function; it is analogous to the usual error term in a traditional cost function. The other component is called an inefficiency component, which in the example here is a half normal.

[8]Measurement error is an even more serious problem for data envelopment analysis, because it may translate into estimated inefficiency. Consider error in measuring factor prices. Suppose, for example, that the hospital actually pays more for its labor than the analyst estimates. The additional payment will bump it off the estimated frontier, and the resulting residual will be attributed to inefficiency. Conversely, if the hospital pays less than the analyst estimates, the frontier will be estimated too optimistically, and other hospitals will appear to be inefficient.

hospital-specific rents in the factor price structure. In practice, wage adjusters for the Medicare Prospective Payment System (PPS) are at the level of the metropolitan area, but it is unlikely that all hospitals in the metropolitan area are in the same labor market or that there are no equalizing differentials within the labor market (it is often argued, for example, that central city hospitals need to pay more to compensate for a higher risk of crime, and a hospital in a small town 50 miles from the central city is unlikely to pay the same wages for a given quality of labor). Thus, the regulator must determine whether wages at a specific hospital that are above average (or above the minimum) are inefficient or equalizing.

6. Error in measuring quality poses an analogous analytic problem to error in measuring factor prices—but probably a more important practical one because of the likelihood of greater measurement error. In my view Zuckerman et al. have made the most serious attempt yet to adjust for quality, but even their effort raises issues with respect both to what is included and what is excluded. Taking first what is included, Zuckerman et al. use the Health Care Financing Administration (HCFA) 30-day mortality variable, but this is well known to be affected by imperfect adjustment for case mix variation across hospitals. (Indeed, HCFA has now suspended its annual release of this measure.)

   Similarly, they use the Medicare Case Mix Index to adjust for case mix variation. Because this is a common practice, it is worth detailing several problems that arise. First, the index applies only to Medicare patients and is imperfectly correlated with case mix variation for those under 65. Second, the index is now based on charges by DRG, whereas previously it was based on accounting cost. Using 1984 data, in 28 percent of cases the cost-based and charge-based weights differed by more than 5 percent (Rogowski and Byrne 1990), suggesting nontrivial error in past weights, current weights, or both. Third, such measures do not account for within-DRG case mix variation.

Most important, the variables included to adjust for quality appear to be far from a comprehensive measure of it. Basing reimbursement on an incomplete quality adjustment raises two further issues beyond those of bias:

- If the included variables are used in a reimbursement formula, hospitals will focus on those aspects of quality for which they are rewarded (or penalized) to the neglect of those for which they are not accountable. Just as in the theory of the second best, where satisfying one marginal condition while leaving others free does not necessarily improve welfare, so too coming

closer to the optimum on one dimension of quality will not necessarily improve welfare.[9] The analogy is the (perhaps apocryphal) manager of the Soviet nail factory. When rewarded by pounds per month of nails produced, the manager produced only large, heavy nails. When rewarded by number of nails produced per month, the manager produced only small, light nails. Soviet carpenters were generally frustrated.

- Hospitals may be able to improve their observed performance through patient selection, assuming the hospital and its medical staff know more than the regulator about the likely outcome. For example, if hospitals are rewarded on the basis of mortality rates that are imperfectly adjusted (because the hospital observes data that the adjuster does not), there is an incentive to not accept (or to transfer) patients whom the hospital knows will make it look bad, e.g., those who have a greater risk of dying than the adjustment formula predicts.

How Much Information Is There?

Over and above the problems with any specific study, there may be something of a Heisenberg principle in estimating hospital cost or production functions; that is, such a function may in fact be for practical purposes underidentified (Breyer 1987). First, and in some ways most important, hospital output is multidimensional. There are 475 DRGs, but a number of DRGs are themselves aggregations of nonhomogeneous cases, so that the true number of "product lines" is considerably greater. These product lines may be produced with varying quality, both technical quality and amenity levels. In addition to inpatient services, hospitals also produce outpatient services, which, because of common or joint costs, must be considered in any production function or cost function estimation. Increasingly they produce post–acute care services (e.g., home health care), and some even engage in unrelated lines of business, whose costs may be difficult to disentangle.

The analyst is attempting to develop a cost (or price) for each product line and each quality level. Because not all hospitals produce all product lines, the analyst must consider economies of scope. Naturally economies of scale also need to be allowed for; all this clearly places additional demands on the data.

---

[9]For an analogous finding see Dranove and Satterthwaite (1992). They show that when consumers select on price and quality, and information is imperfect about both, improving information on only one dimension does not necessarily improve welfare.

A natural method for accommodating economies of scope and scale, as well as uncertainty about true functional form, is to use a so-called flexible functional form (Breyer 1987, Jorgenson 1986, Lau, 1986). Such a form requires estimating a parameter for each output and each input price, as well as a parameter for each cross-product term in outputs and input prices. Suppose, to oversimplify the problem, there is only one input price, a wage index. Suppose to further oversimplify there are not 475 outputs, but only 100, including outpatient department and emergency room outputs, and that there is no quality variation to consider. A flexible functional form requires estimating 1 + 100 + [(101)(100)/2] or 5,151 parameters.[10] Roughly speaking, the number of parameters rises with the square of the sum of inputs and outputs.

Alas, in any cross section one has only roughly 5,000 hospitals available because that is the number of hospitals in the country. Time series variation is not likely to provide much additional useful variation, in part because hospitals are reasonably stable enterprises from year to year with respect to inputs and outputs (e.g., the size of the nursing staff probably will show only modest variation) and in part because some of the variation over time stems from technological change, so that one cannot presume that the structure being estimated is constant. Moreover, even in a cross section certain environmental variables, such as the state regulatory environment, would normally be presumed to affect cost, requiring the estimation of additional parameters.

Technological change over time raises an additional issue beyond that of precluding the analyst from pooling data across several years. In real time, estimated cost or production functions are likely to be from data that are several years old. It may be, therefore, that the calculations of any estimated frontier function are no longer relevant to the current rate-setting situation.

Analysts have, of course, estimated cost functions. To do so, they have imposed more structure on the problem than is imposed by a flexible functional form and have largely abandoned the pure flexible functional forms. One widely cited cost function estimate is that of Granneman et al. (1986), who use what they dubbed a "hybrid" function (a hybrid between methods that simply use average cost per patient day or stay with numerous regressors and structural cost models with flexible functional forms).[11] To decrease the number of parameters being

---

[10] In its simplest form (linear in logarithms and their cross products), it also requires nonzero levels of all outputs for each observation, which will not hold in reality if the inpatient side is at all disaggregated (e.g., not all hospitals have intensive care units or treat all diagnoses), nor do all hospitals have outpatient departments. Generalized translog functions formally do not need this assumption but often yield unstable results with input levels near zero.

[11] A similar approach was used by Hadley and Zuckerman (1991).

estimated, these authors reduced the number of outputs to five, the number of input prices to four, did not take cross products of input prices and outputs (this imposes homotheticity), and did not account for all cross products within the output vector. They therefore estimated only 64 parameters; it seems clear, however, that such output measures are too aggregate to use to identify efficient hospitals (e.g., Granneman et al. aggregate all acute care days and discharges, all subacute care days and discharges, and all intensive care days and discharges).

Nonetheless, with their 64 explanatory variables Granneman et al. explained around 93 percent of the variation in cost across 867 hospitals. This suggests that one will not be able to precisely estimate very many more parameters (say ten times as many); yet that is what identification of efficient hospitals with realistic assumptions about the dimensionality of inputs and outputs would seem to require.[12]

Zuckerman et al. (1993) aggregate to five outputs (inpatient admissions and postadmission inpatient days for Medicare and non-Medicare patients considered separately, and outpatient visits for all payers); they consider two input prices and estimate 74 parameters using a sample of 1,271 hospitals observed over a seven-year period; their R^2 is around 0.97, again implying that a substantially greater level of disaggregation with inputs or outputs is not likely to succeed.

In addition to the problems already described, two other measurement difficulties, both of which are reasonably well known, should be mentioned. The first, measurement of capital inputs, is almost always a problem in cost and production function estimation. The studies in the literature deal with it in varying ways; many cost function studies simply omit capital inputs and treat

---

[12]An alternative line of argument is that reliable measurement of many parameters is not necessary because these high correlations show that fewer parameters are necessary to explain hospital costs and that accuracy in estimating predicted cost, not accuracy in the estimation of structural parameters, is what the regulator needs. This argument neglects two issues.

First, one needs the strong assumption that not only do relatively few dimensions summarize hospitals (because of collinearity among dimensions), but that if reimbursement methods were changed, behavior would not change. Put another way, the assumption is that the collinearity among the various dimensions reflects technology or something else outside the hospital's control. This seems dubious. One might consider the following analogous argument. One can probably reduce the number of DRGs (i.e., aggregate cases into fewer, more heterogenous DRGs) by a substantial factor (half?) without much changing the R^2 in a regression of hospital costs on DRGs. One would expect maximizing hospitals, however, to exploit the greater heterogeneity in the reduced number of DRGs. Moreover, given the changes occurring in the delivery system (e.g., hospitals becoming "full-service" hospitals to compete for the business of managed-care plans, hospitals merging or closing), the observed historical collinearity may be breaking down.

Second, in any actual reimbursement scheme, one must reimburse hospitals, and there will be deviant hospitals that do vary in these characteristics; specialty hospitals come readily to mind. Thus, estimation of a larger number of structural parameters is in fact necessary. The example of hospitals exempt from PPS serves to remind that a regulator must deal with all hospitals.

the estimated function as conditional upon existing capital stock. Those that include it almost invariably use an accounting definition of capital; that is, the conceptually correct foregone interest plus the change in the value of the capital stock (depreciation plus capital gain or loss) is not used; rather an accounting approximation is used. Often this does not attribute an opportunity cost to donated capital and uses an assumed economic life for the asset, which may or may not correspond reasonably with reality.

The second input that is often not included is that of physicians.[13] It is clear that physician inputs should be included if the relevant output is deemed to be treatment of some disease; that is, if hospital and physician services are both intermediate goods. Even if hospital services are considered as final goods, however, physician inputs will affect measured quality (e.g., misordering by physicians would be one cause of the drug incidents used by Zuckerman et al. as a measure of quality), and physician inputs may well substitute for conventional hospital inputs. One example of substitution occurs in a teaching-hospital setting where resident time substitutes for both physician and nurse time. Analogous to physician inputs, inputs of contract nurses are often omitted from empirical studies.

Is the Hospital Too Aggregate a Unit?

The task of developing a method to measure costs at efficient hospitals is framed as though each hospital were equally efficient at everything it does, but this is unlikely to be the case. An example is mortality outcomes for various procedures; the hospital may be high on one procedure and low on another. Although some observed variation is random, some surely reflects true variation, such as the capabilities of the various surgeons and nurses involved in performing different procedures. (This example again illustrates the need to consider physician inputs in estimating production or cost functions for hospitals.) Another example is departmental inefficiency; if a particular department such as the laboratory or the pharmacy is badly managed, it will affect patients differently according to the degree their treatment draws on that department.

Thus, if one wished to identify units that were producing efficiently and reimburse their costs, which would seem in the spirit of the Boren Amendment, the hospital might be too aggregate an enterprise. However, analysis at the department or case level to find the best performing departments poses severe

[13]Some exceptions are Jensen and Morrisey (1986a, 1986b) and Custer et al. (1990).

problems. Allocation of joint costs becomes an issue, and the treatment of economies of scope becomes more important.

Furthermore, if one found the minimum for each department or subunit and used that figure to reimburse, the result could well be to drive all actual hospitals out of business (or cause them to adjust their quality) rather than become more efficient; put another way, an actual market would not necessarily lead to such reimbursement rates. Consider commercial banking in major metropolitan areas, which is usually thought to be reasonably competitive. It seems unlikely that all departments in all large commercial banks are equally efficient.

4. Conclusions

One response to the foregoing discussion may be that it is asking for utopia when utopia is not necessary. The practical question is whether budget or rate regulation based on estimated frontier cost or production functions is an improvement over an alternative approach of making an incremental change from the status quo and observing the consequences.

One indication is provided by practice in other countries. As far as I am aware, no country uses an explicit cost or production function to determine a rate, much less to determine the cost of an efficient hospital. And while I cannot specify the comparative degree of welfare loss in using an inevitably misestimated frontier in a reimbursement formula relative to the approach of simply making an incremental change, the foregoing criticisms of the frontier technique lead me to conclude that the degree of misestimation and hence the welfare loss could be serious. Indeed, I believe that if frontier cost methods were actually used to set rates welfare losses are likely to increase. One might be able to use the technique as a screen to identify outlier hospitals, but it is not clear that even upon closer inspection, one could readily distinguish inefficiency from "justified" costs. If this conclusion is accepted, does the economics literature have anything to say about how a regulator should reimburse hospitals?

The appendix contains a selected review of the literature on the reimbursement of regulated firms when the regulator is uncertain about costs. Although qualified, the conclusion from that literature is that reimbursement should not be wholly prospective, but rather partially prospective. The current Medicare PPS is wholly prospective; an example of a partially prospective system would be a linear combination of the current PPS and the old cost-based system.[1]

The conclusion that reimbursement should not be wholly prospective emerges from a wide variety of models, which emphasize different features about hospital

[1] I argued above that costs are imperfectly measured, and it may seem inconsistent to now argue that reimbursement should be based upon these imperfectly measured costs. One should distinguish two issues. First, total costs are imperfectly measured because of the omission of certain costs, notably physician and contract nurse costs, as well as the omission of capital costs (or the failure to use economic costs of capital). These problems are potentially remediable, although my judgment is that there are gains to be had from using hospital-specific total costs in the reimbursement formula even if they are not remedied. Second, the cost (average or incremental) of a particular product is imperfectly measured. This is not an issue in using institution-wide costs in the reimbursement formula.

care. The appendix begins by considering Shleifer's (1985) model, which reaches the conclusion that payment should be entirely prospective. The appendix then considers models that modify Shleifer's assumptions in the direction of greater realism; these modifications lead to a partially prospective and partially cost-based system. For simplicity I have only considered linear combinations of prospective and cost-based systems; although one might theoretically do better with a nonlinear system, even linear systems strain the state of the art.

Shleifer's (1985) model assumes that a large number of profit-maximizing firms produce a homogeneous and observable (by the regulator) product; in such a case, reimbursing at the average marginal cost (combined with a lump sum transfer if needed to satisfy a break-even constraint), will lead firms to produce efficiently. Intuitively, firms face a price that acting singly they cannot influence, and they keep any residual savings; hence, they equate at the margin saving a dollar of production cost with the cost of the effort needed to save that dollar; the relevant first-order condition for efficiency in production is satisfied.

Shleifer's model seems suitable for a regulated utility such as electric power, in which the product is homogeneous, readily measured, and unidimensional (e.g., kilowatt hours). The thrust of the foregoing analysis suggests that it is ill suited to hospitals, despite Shleifer's use of the PPS as a real-world example of his model. I note in passing that Shleifer's assumptions imply that all observed cost variation among hospitals is attributable to inefficiency, an implication shared with DEA.

The health economics literature has developed models that relax one or another of Shleifer's assumptions. In so doing, the conclusion that reimbursement should be wholly prospective is also modified. The foregoing material emphasized that the product across hospitals is not unidimensional but heterogeneous and multidimensional; moreover, it is not fully observed by the regulator. Heterogeneity in the product of different hospitals extends to treating patients of varying severity or complexity.

One class of models in the health economics literature emphasizes that the heterogeneity in patient mix across hospitals, as well as variation in efficiency, affects observed variation in hospital costs. Assuming the regulator desires to reimburse for the unobserved variation in patient heterogeneity (i.e., the regulator's desired price should reflect the variation in patients across hospitals), one can minimize the expected mean square error of pricing errors by using a linear combination of the average price of a large number of hospitals and the hospital's own costs (Pope, 1990; Goodall, 1990; Keeler, 1990). In a simple model the weight on the hospital-specific costs is simply the proportion of variation in

the residual of the cost function that is attributable to the unobserved costs that the regulator considers reimbursable. In the PPS context, Keeler (1990) estimated this weight to be at least 15–30 percent, assuming that hospital efficiency is not affected by changing to this reimbursement method. To the degree efficiency is affected by basing reimbursement on observed costs, the optimal weight on hospital costs is lower and it cannot be ruled out that reimbursement should be wholly prospective.

A second class of models comes from Ellis and McGuire (1986, 1990), who relax Shleifer's assumption of profit maximization. They construct a model in which treatment is determined by a physician, who is assumed to act as an agent for both the patient and the hospital. What the physician seeks to maximize is critical in determining the outcome. Ellis and McGuire assume that physicians' utility functions have as arguments benefits to patients and profits to the hospital. If the physician weighs a dollar of patient benefits equally with a dollar of hospital profits (at the margin), fully prospective reimbursement is optimal. If the physician, however, gives greater weight to hospital profits, the regulator can offset this by decreasing the share of residual profit that the hospital can keep by introducing an element of cost reimbursement.

The problem of contracting with firms when the regulator does not know the firm's cost has also been discussed in the general industrial organization literature by Laffont and Tirole (1986) and Laffont (1987).[2] Laffont adds to Shleifer's model the element of heterogeneity in cost across firms, a portion of which the government wishes to reimburse, as in the Pope model. Further, unlike any of the above models, the taxes that finance the government's expenditure involve an efficiency loss; thus, there is a penalty to reimbursing inefficient hospitals. [This loss tends generally to be overlooked in the health economics literature, but see Ballard and Goddeeris (1993).] In Laffont's model the government wishes to keep all firms in business, although this can readily be modified to specify a cost above which a firm would not be kept in business.[3]

In Laffont's model the firm makes a bid. Thus, this type of model appears relevant to states that selectively contract in their Medicaid programs (and also seems relevant to choices by health maintenance organizations and PPOs that do not necessarily contract with all hospitals).

---

[2]See also the survey by Baron (1989).

[3]Whether the government should want to keep all hospitals in business can certainly be questioned, but it arguably does represent actual government behavior with respect to hospitals.

Laffont structures the game between the regulator and the firm as follows: The firm announces a bid. The government accepts the bid, at least if it is below a certain level, but varies the terms of the contract the firm receives depending on the bid. The contract, however, is always a linear combination of a fixed-price and a cost-reimbursement contract. The lower the bid, the more the contract will resemble a fixed-price contract and conversely. Laffont assumes—and this is important—that the government knows the range of firms' (hospitals') costs if they exert the socially optimal amount of effort to be efficient, though it does not know the cost of any particular firm or hospital.

The intuition of this model is that the government's optimal strategy is to let the firm keep savings from holding costs below a fixed price, the lower the firm's fixed price is. The exact mathematical results of Laffont's model depend on a number of strong assumptions, but the intuition that a pure fixed price or prospective system is not optimal would appear to go through on weaker assumptions.

In general, one can view accounting for hospital-specific costs as a compromise in an inevitable trade-off between selection behavior (seeking patients for whom revenue exceeds expected cost) and efficiency. For more than two decades health economists have talked about the trade-off between moral hazard and risk aversion (Zeckhauser 1970). More complete insurance provides more risk protection at the price of greater consumption of services that are valued at less than their social cost. The burden of the argument here is that there is an analogous trade-off between selection and efficiency.

Fixed-price contracts (fully prospective reimbursement) give the maximum incentive for efficient production. But the technology for setting the fixed price at the expected cost of treating a given patient efficiently does not exist. Thus, the true expected cost for any given patient, about which the hospital is likely to have more information than the regulator, may exceed or fall short of reimbursement. Fixed-price contracts thus give the maximum incentive to select against patients whose costs exceed reimbursement, whereas traditional cost-based reimbursement gives none.

Selection with a fixed price and heterogeneous product can be observed in a supermarket that is selling produce at a fixed price per pound. Because shoppers inspect the produce before they buy, at the end of the day it is the bruised apples and the overripe bananas that they have left on the produce stand.

A Multiple Payer System

The analysis to this point has not considered a multipayer system. In such a system the allocation of joint costs among payers is arbitrary. They must be covered or the hospital will not break even, but there is an incentive for every payer to pay marginal or incremental cost. The institution has an incentive to shift these costs to the payer with the less elastic demand curve, which tends to produce cost shifting to programs such as workers' compensation.[4]

Medicaid rates have historically been among the lowest of all payers. If rates paid by one payer (Medicaid) fall sufficiently relative to private rates and cross subsidization is not possible, one could see hospitals specialize by payer and exhibit different cost levels.[5] (If cross subsidization is possible, outcomes will depend upon hospitals' willingness to engage in it.) Assuming specialization or segregation is not desired for reasons of equity, the Health Care Financing Administration might monitor whether the concentration of Medicaid (or Medicare) patients is changing over time as a test of whether Medicaid reimbursement has fallen too far. In doing so, it would be necessary to control for any changes in the geographic distribution of Medicaid or Medicare eligibles (e.g., recent eligibility expansions in Medicaid may well have increased the geographic dispersion of eligibles within local markets).

Increasing concentration may reflect a greater fall in rates than is desired (indeed segregation may lead to higher social costs if more hospitals are needed to specialize in patients of different payers and fixed costs are important).

---

[4]Standard Ramsey pricing is not really an issue here, since these are supply prices (prices paid providers by the insurer), not prices to patients. A modification of standard Ramsey pricing would be relevant if payment above marginal cost generated varying degrees of supply-induced demand.

[5]A simple model is as follows. Let the cost function for a given quality level be quadratic in a measure of output (such as case mix adjusted discharges) with an additive fixed cost term. Let there be two quality levels with two quadratic functions if production is separate. Let revenue be a constant (case-mix adjusted) figure per discharge, with each of two payers paying different constants and each payer having finite numbers of patients. The marginal revenue function if care is produced in the same hospital for both payers is therefore a discontinuous step function with a break at the number of patients from the higher-paying payer. If care for the two payers is produced in different hospitals, the marginal revenue function is simply a constant until the number of patients is exhausted.

Depending on the magnitude of the coefficients in the two cost functions, it may cost less to produce care of the same quality for both payers at the same hospitals or have twice as many hospitals specializing in caring for each payer's patients. Assume it costs society less to produce care for both payers together at the higher-quality level than separately at different quality levels (for sufficiently high fixed costs and sufficiently low-cost differences in quality levels, this will be true). If reimbursement exceeds marginal cost for high-quality patients over the range of low-paying patients, hospitals will clearly not specialize. If, however, reimbursement for the low-cost payer (Medicaid) falls below marginal cost at the high-quality level and high-quality hospitals do not cross subsidize Medicaid patients, there will be an opportunity for low-quality hospitals that specialize in Medicaid patients to enter (assuming that Medicaid reimbursement exceeds [the lower] marginal cost at the lower-quality level by a sufficient amount to cover fixed costs).

Monitoring the concentration of patients is simpler by far than estimating marginal costs.

From a social point of view, there is the further consideration that various payers impose different distortions in the process of raising funds. There are several estimates of the inefficiency in raising the general revenues that finance the Medicaid program (Ballard and Fullerton 1992). Unfortunately, there is a considerable range in these estimates, but a middle-of-the-road estimate might be around 30 percent.

The other major hospital payers, Medicare and private insurance, raise funds from payroll taxes and from employer- and employee-paid premiums. Raising revenues in this manner also causes inefficiencies, although their magnitude is not well documented. One could possibly rationalize lower Medicaid rates as a response to a higher perceived shadow price on raising general revenues, although this is highly speculative. Any complete calculation would also have to consider the social cost of any diminished access or lower quality care that Medicaid beneficiaries received as a result of the lower rates, as well as any incentives to obtain Medicaid eligibility if rates increased.

Recommendations

To this point I have argued the following.

1. The direct approach of trying to measure the costs of an efficient hospital has serious problems. Even the measurement of cost is imperfect; capital inputs are not well measured, nor are physician inputs. More fundamentally, the unexplained variation in costs across hospitals stems both from reasons that are to some degree beyond the control of hospitals (e.g., unobserved case mix variation) and from reasons that are not (e.g., varying quality of management and varying level of effort by management) in ways that cannot be empirically dissected in any ready fashion.

   Furthermore, accounting for economies of scale and scope as well as quality variation and input price variation would appear to strain the available data; a fully flexible functional form with 100 outputs and inputs requires approximately as many free parameters as there are hospitals. Additionally, medical care technology is constantly changing; by the time one had estimated this function, it would be in part obsolete.

2. Even if the cost (or production) function were well understood, the regulator must choose a quality level. That is, by choosing a reimbursement rate, the regulator is choosing a point on the function. Put another way, because costs

vary with quality along an efficient frontier, the legislative injunction to reimburse at the cost of an efficient facility is not sufficient to determine a rate.

3. The maximum incentive for hospital management to be efficient is given by a lump-sum payment for a patient (i.e., a payment method that is independent of management actions under which the hospital can keep the residual, e.g., PPS). Unfortunately, if the payment is the same for all patients in a class but those patients are not homogeneous, the hospital has an incentive to select those patients who are profitable (at the margin) and "dump" those patients who are not. If all hospitals have such incentives, the unprofitable patients will have access problems. There is thus a trade-off in determining a basis of payment for hospitals between efficiency and access (selection).

   An analogous problem arises if hospitals are rewarded or penalized for quality (e.g., mortality rates), but the adjustment for how case mix variation affects the measures of quality is imperfect and the hospital's information is better. Hospitals will then have an incentive to shun patients who will make their quality look bad.

4. The trade-off between efficiency and selection is embodied in how much weight the reimbursement formula places on a group average (PPS places all the weight on the group average) and how much it places on hospital-specific costs; i.e., the degree of prospectivity.

The literature suggests that perhaps 15–30 percent of the weight should be placed on hospital-specific costs, but this is under special assumptions. Desirable weights will almost surely have to be determined empirically. Experimentation with "partial capitation" would be desirable.

Even if one accepts, however, that reimbursement should take some cognizance of the hospital's costs, that is not sufficient to determine the level of payment nor, in a multipayer system, the difference, if any, in payments among payers. How then should this be determined?

If one believes that a great deal of uncertainty attaches to any empirical estimate of the frontier and therefore rejects basing reimbursement on such an estimate, one is left with a traditional strategy of "incremental groping"; that is, making modest changes from the status quo and monitoring the effects of those changes. Such a strategy is similar to what has recently been followed with Medicare hospital reimbursement; year-to-year changes have been made in PPS rates with some adjustments in structure based in part on experience or feedback.

Such a procedure is admittedly theoretically unsatisfying. Some incremental actions might be taken that cause irreversible changes. There is no guarantee one will converge to a global optimum if there are multiple optima. Nonetheless, from a pragmatic point of view, this may well be the best that can be done at the present time.

This strategy places a high value on the goodness of the monitoring system. If feedback is late in coming or yields a false picture, welfare losses under this system will clearly be greater. (The same timing issue, however, applies to basing reimbursement on an estimated frontier; if the frontier is estimated with out-of-date data, the resulting rates could be far from optimal.) At a minimum, a monitoring system should include mortality and readmission rates, risk adjusted to the degree possible, as well as patient surveys to elicit functional outcomes and patient satisfaction. It should also include a sample of chart reviews to determine the technical quality of care and the adverse event (iatrogenic injury) rate. Possibly cost or production function estimation could be used to target particular hospitals for review, but it would remain to be shown that more intensive review could satisfactorily distinguish inefficiency from "justified" costs.

In designing a monitoring system, one might examine Kaiser-Permanente or other private systems that operate hospitals and are thought to have incentives to be efficient because of reimbursement from a capitation rate. What indices do they use internally to judge which of their hospitals is running efficiently? What do they measure, if anything, to explain cost or quality variation across their hospitals, or more generally to determine the performance of their hospital managers?

A somewhat separable concern in a multipayer system is whether a given payer's rates are "adequate," and in particular in the current system, whether Medicaid rates are too low. The answer to these questions depends upon the magnitude of several factors: (1) any differential in access to services that is induced by the payment differential, (2) any difference in quality of care that is received as a result of the differential, (3) the distortions induced as a result of raising general revenues to support Medicaid relative to distortions from raising monies from other sources (especially payroll taxes and employer-paid premiums). To the degree that there are differential payments among states, the consequences of this variation could be examined.[6] Changes in the differential in access can be

[6]Decker (1993) has recently followed such a strategy for Medicaid physician payments. She finds that Medicaid beneficiaries in states that pay higher fees are no more likely to visit a physician in a year, but the physician is more likely to be a specialist and will spend more time in the visit.

partially monitored by examining whether concentration of Medicaid recipients by provider is changing (controlling for location of residence).

Adopting an incremental or empirical strategy for rate setting does not obviate the need to address the issues raised in the discussion of cost and production function estimation. There is still a need to choose, in effect, a level (or distribution) of quality, if quality is a function of rates or budgets, as one presumes it is on the frontier. There is a need to value the various output dimensions; a change in the level of rates may improve one dimension of output and worsen another. Monitoring is inevitably incomplete, and even incremental changes may reward hospitals for certain observable dimensions and not for other unobservable dimensions, leading to possible distortions. The major advantage in following this procedure rather than setting rates at estimated frontier values is that incremental change protects against the possibility of large losses from what subsequently turn out to be estimation mistakes. And incremental change should be easier to implement.

Appendix

A. The Basis of Reimbursing Health Care Providers

Several recent articles in the health economics literature (Ellis and McGuire, 1986, 1990; Goodall, 1990; Pope, 1990; Keeler, 1990; Selden, 1990; Newhouse, 1991; and Siegel et al., 1992) have contributed to a deeper understanding of a debate that has gone on for many years: How should health care providers be paid, by fee for service, by capitation, or by salary or "budget"? And if fee for service, what defines a service? In this appendix, I try to review and synthesize this literature and draw some parallels between it and a related literature in industrial organization on procurement.

The traditional view of this question, still often heard in policy debates, was harsh on the fee-for-service system, suggesting that it provided incentives for overservice.[1] The more recent view suggests that all systems have drawbacks, and that a mixed system, for example part capitation and part fee for service, may offer advantages over a pure system of any type.

My main point in this appendix is thus to suggest the consideration of mixed, or two or more part pricing (two or more bases of pricing), because of various imperfections in the market. These imperfections include the presence of insurance (or public provision of services) and informational asymmetries between producers, payers, and consumers.

Before proceeding, I want to digress somewhat to clarify the term "budget" and why I placed it in quotation marks in the opening paragraph. I have included the term in the list because it apparently describes reality. Hospitals, for example, may literally receive a budget (as in Canada); one might also regard a salaried physician who is an employee of an organization delivering services as having a budget for his or her time. Thus, by budget I mean a lump sum that is set with the intent of paying for some number of services or hours but is in the short run independent of workload. The short-run caveat, however, is why I placed

[1]The criticism is hardly new, as is evidenced by the well-known quote from Shaw: "That any sane nation, having observed that you could provide for the supply of bread by giving bakers a pecuniary interest in baking for you, should go on to give a surgeon a pecuniary interest in cutting off your leg, is enough to make one despair of political humanity." George Bernard Shaw, *The Doctor's Dilemma*, 1911.

budget in quotation marks; in the long run there may be some relationship between budget and workload.

Such a relationship between budget and workload will often be implicit, although it may be explicit. Examples of the latter are Veterans Health Administration hospitals in the United States, whose budgets at times have been adjusted by the DRG system to reflect the case mix and volume of cases treated in each hospital. In the case of a salaried physician, the issue is what determines future salary; are salary increases or bonuses tied to performance, including the workload processed?[2] To the degree that future budgets or salaries are some function, perhaps implicit, of the organization's or person's workload, a budget in the long run has elements of fee for service or capitation.

Traditional Fee-for-Service Payment

I begin with the traditional fee-for-service system. If consumers were well informed (including being able to monitor the work the physician does), if the market were competitive, and if there were no insurance, a disaggregated fee-for-service system would be the expected payment system and would have the standard claims to optimality.[3] Any other system would have some aspects of a tie-in sale or bundling of services that could not yield an efficiency gain. That fee for service would be the expected system follows from tie-in sales not being possible in a perfectly competitive market—and comports with the observed fact that fee for service is the system of payment traditionally used to pay for private health care services in less-developed countries where there is little insurance. Its claim to optimality is discussed in Pauly (1980) but is at bottom the standard claim to optimality in a perfectly competitive market.

One could ask about the ability to generalize this claim to professions other than physicians; no one would seriously propose, for example, that the President of the United States be paid on a fee-for-service basis. I have not made the effort

---

[2]One potential advantage of salary over fee for service or capitation for physicians is if quality is more accurately rewarded because another physician is evaluating performance and the other physician is a better judge of quality than the consumer/patient. On the other hand, there are three potential drawbacks to salaried employment with a physician supervisor assessing performance: there may be incentives to shirk; only the consumer may be able to judge certain aspects of quality (though a manager could sample consumer opinion); and there may be insufficient reward for merit. I deal with some of these issues below, but the point to make here is that a manager, most likely a physician, who is setting the salary of an employed physician may well use different criteria than a patient would use.

[3]I ignore transactions costs of these various systems, which in this context means I assume that any additional cost to bill for each item is negligible. This may not always be accurate, and in some cases global fees may economize on costs, but I wish to focus on other issues. I also ignore issues of risk aversion; for example, a global fee for all prenatal care relative to a fee for each visit shifts risk from the patient to the provider; again, this is not the issue on which I wish to focus.

here to explore conditions for an optimal payment system, but note that a principal's ability to monitor effort and attitudes toward risk will play a role (Stiglitz, 1974).

Traditional health insurance introduces an obvious problem for optimality because it is usually written as a subsidy for services actually used rather than as a lump sum transfer conditional on some exogenous state of the world. If income effects were small and if there were no other distortions, a standard argument for overuse follows (Pauly, 1968); however, income effects are not necessarily small and in fact there are other distortions.

Two of those distortions are the price paid by the insurance company, which may not be a perfectly competitive price, and the information available to the consumer, which may be less than is available to the provider.[4]

The Price Paid by the Insurance Company

The insurer clearly cannot agree to reimburse whatever the provider of services bills because that, in principle, is an unlimited amount. One can achieve a theoretical limit on the provider's billed price through coinsurance, that is the consumer's paying a percentage of the bill (Frech and Ginsburg, 1975); in that case, the billed price is limited by the consumer's willingness to pay the coinsurance. Many actual insurance policies, however, have no coinsurance at the margin, and there may be bad debt (consumer refusal or inability to pay) even in those that do. Thus, the assumption that coinsurance induces competitive pricing is one I do not wish to make.[5]

Instead, I assume that a price to be paid the provider is set by the insurer, which may be the government in the context of either public insurance or public production.[6] The price reimbursed by the insurer need not be the price that would have been observed in a competitive market with no insurance, even

---

[4]There is a third potential distortion from externalities, that is, person A's willingness to pay for the treatment of person B. I mostly ignore this distortion in order to focus on the basis of reimbursement.

[5]Even if there were, say, a 10 or 20 percent coinsurance rate, there may well be a poor approximation to a competitive outcome because of the reduced incentives to search for the most efficient (lowest price for a given product) supplier.

[6]In principle, unit price differences among different providers could be reflected in a consumer's premium, and indeed, the development of PPOs can be viewed as an attempt to achieve this. However, it is not always the case that price differences among providers are reflected in what the consumer pays, and in any event we have not addressed what the unit of service is that is being priced.

ignoring the possibility of large income effects in an uninsured market.[7] The Medicare program, for example, which reimburses for most cataract operations, may not know what the price of a cataract operation would be in a perfectly competitive market.[8] Turning this question around, there is a question of how the insurer or the state should reimburse providers so as to minimize rents.

Agency

A second distortion comes from the provider's role as the patient's agent. For concreteness one could think of the provider as a physician, but it could also be an organization such as a health maintenance organization. Some choices will be made by the provider—or equivalently for my purposes, made by the consumer relying on information supplied by the provider—so the question arises, does the provider have incentives to act (or to supply information) in his or her own interests and in ways that are not necessarily perfectly congruent with the patient's interests if the patient knew as much as the provider? This issue has been the subject of a vast and well-known debate in health economics over supplier-induced demand. My purpose here is not to add to that debate, but rather to enquire into the implications of the information asymmetry for pricing.

Although I began this appendix by discussing fee for service and capitation, theory often follows from events, and much of the theory I will discuss was stimulated by the enactment of the American PPS in 1983. The intent of the PPS was to achieve greater efficiency. The problem that the literature sought to illuminate was whether or to what degree patient- or hospital-specific costs of treatment should be considered in reimbursement, or whether the government should simply pay the same amount to each hospital (adjusted for local factor price differences) for every patient in the same DRG.

This prospective system can be contrasted with the prior system, which can be regarded as a variant of a fee-for-service system—it reimbursed a cost or a "fee" for each drug, for each supply, for each unit of time in the operating room and recovery room, etc. Thus, if the PPS were to be modified to consider patient-specific costs, one could consider that a variant of a fee-for-service system. Although paying a fixed amount for each admission, as PPS does, could be

---

[7]That is, I wish to take the case where amounts spent are small relative to income and so ignore the problem that for expensive treatments there may be little demand in an uninsured market.

[8]It could potentially learn this price through a bidding mechanism, but that would imply that beneficiaries would have to use care provided by the lowest bidder. For some services, in which there is no personal relationship, such as laboratory services, a bidding arrangement seems plausible, but for personal physician services the program has scarcely put any limits on patient choice of physician.

regarded as a "fee per admission," it was not only a more aggregated pricing system than the system it replaced, it did not vary with the services delivered. The limit in aggregating cost is capitation; all services for a defined period of time are paid for with one price.

The Organization of the Appendix

Andrei Shleifer (1985), in a well-known paper on yardstick competition, has described a model under which fully prospective payment methods such as the PPS achieve optimality. I therefore begin with his paper. Although Shleifer uses the Medicare PPS as a real-world analog of his model, in fact his model seems better suited to public utilities, such as those producing electric power, than to hospitals. To demonstrate this I turn to a number of papers from the health economics literature that relax one or another of Shleifer's assumptions and reach different conclusions; in my view these papers make more reasonable assumptions about the hospital market. In any event, the contrast between the models in the health economics literature and Shleifer's shows how violations of Shleifer's assumptions will invalidate the conclusion that fully prospective reimbursement, as in the PPS, represents optimality. I close with some remarks on related papers from outside the health economics literature.

Shleifer: Yardstick Competition and Prospective Payment

Under certain restrictive conditions, a pure prospective system, for example the current PPS, is better than the mixed reimbursement system I consider below. A model that reaches this conclusion was introduced by Andrei Shleifer (1985).

The essence of Shleifer's model is that a regulator sets prices for firms in a fashion similar to the PPS. Firms can invest in cost-reduction efforts; the issue for the regulator is to obtain efficiency in firms' efforts at cost reduction (distributional considerations are ignored). If the regulator simply reimburses costs, as was the case in Medicare prior to PPS, there is no incentive for the firm to engage in cost-reduction efforts. The regulator can, however, set a fixed price, in a sense that can be made precise, which will elicit the optimal cost-reduction effort and hence efficiency.

In the remainder of this subsection I describe Shleifer's model more formally. Shleifer assumes that there are n identical firms—or firms that are identical after adjustment for measurable characteristics. This assumption is in fact the assumption that a portion of the health economics literature seeks to relax, but it

is useful to describe Shleifer's model to show that with this assumption, full prospective payment can be optimal.

The firms in the model can make an investment in cost-reducing technology (say buying a computer to reduce billing costs), but there is no incentive to do this if costs are simply reimbursed. Strictly speaking, with cost reimbursement the firm is indifferent about buying the computer or not, but Shleifer assumes that there is some nonreimbursed effort involved in making the investment, so that in fact the firm or hospital is not indifferent and prefers not to make the investment in a cost-reimbursement world. The assumption about effort is in fact important (and becomes more explicit in Laffont's model discussed below), because if all costs of the investment are monetary (e.g., no nonmonetary costs of entrepreneurial effort), the firm has the proper incentives if it is allowed to keep any share of the profits; the share can be made arbitrarily small.

In Shleifer's model there is an initial constant marginal cost of production c_0, which the firm can reduce to c by spending R(c) on cost-reduction efforts. Thus, $R(c_0) = 0$ and $R'(c) < 0$. Profits of the firm are

$$V = (p - c)q(p) + T - R(c) \quad , \tag{1}$$

where p is unit price, q is quantity sold, and T is a lump-sum transfer made to the firm that is collected through lump-sum taxes. If the regulator has perfect knowledge, the regulator picks c, p, and T to maximize the sum of consumer surplus and firm profit, subject to a breakeven constraint $(V \geq 0)$.[9] The solution to this problem is to choose c such that

$$-R'(c^*) = q(p^*), \tag{2}$$

$$\text{and } p^* = c^* \text{ and } R(c^*) = T^*.$$

Equation (2) is simply the condition for cost minimization for producing output level q; this can be seen by differentiating Equation (1) with respect to c and setting the result equal to zero. Intuitively, this requires that the hospital invest in cost-reducing technology until its incremental investment cost $R'(c)dc$ (or $(dR/dc)dc = dR$) equals the incremental savings in production cost, $-q(dc)$. The condition $p^* = c^*$ is simply price equals marginal cost, and the condition $R = T$ is required to have the firm break if $p = c$ [see Equation (1)].

[9]Those who believe that externalities are important will note that any definition of consumer surplus should include these externalities. Shleifer's problem, however, was to obtain efficient production; for this problem his formulation seems adequate. In effect, externalities must be addressed through subsidies on the demand side.

Intuitively, if one accounts for investment in cost-reducing technology, there is always declining average cost, because the variable cost of production c is constant and there is a fixed cost equal to R. Hence, if one prices at marginal cost, a transfer equal to R is required to break even. As a result, if an omniscient regulator were to announce the prices p^* and T^*, a first-best outcome would be achieved. By contrast, if the regulator announces that price p will equal whatever c the firm chooses and price T will equal whatever R the firm chooses (cost-of-service regulation or cost reimbursement), the first-best outcome will not be achieved since the firm will choose R equal to zero and c will equal c_0.

A regulator, of course, is not omniscient and hence does not know c^* and R^*. Shleifer's contribution is to show that the first best can be achieved under a weaker condition; the regulator sets p to equal the average observed costs among other identical firms in the market and R to equal the average investment R among other identical firms in the market.[10] If the number of identical firms is large, there is little loss from simply setting a price equal to the average across all firms. Shleifer also considers the case in which no transfers are allowed; he shows that prospective payment based on observed average cost is second best. In effect, Shleifer's insight is that a fixed pricing rule based on observable data can lead firms to efficient production. Indeed, Shleifer cites the PPS as a real-world analog to the pricing system in his model, in that the PPS price is set equal to the average of all firm's prices for a particular DRG (with some additional adjusters that are not patient specific).

Pope, Goodall, and Keeler: Unobserved Patient Heterogeneity and a Mixed Reimbursement System

A class of models has been developed by health economists that emphasizes unobserved heterogeneity among firms (hospitals) because of unobserved patient heterogeneity (Pope, 1990; Goodall, 1990; and Keeler, 1990). Thus, the assumption of identical firms made by Shleifer is relaxed in these models. As we shall see, relaxing this assumption changes the conclusion; an optimal reimbursement system in these models is a mixed system that is partly a payment based on the average cost of many hospitals (in treating a particular case), and is partly hospital or patient specific. Thus, this system differs from the uniform payment system represented by PPS, which is purely based on the average cost of many hospitals. For conceptual purposes, one might think of the

[10] Actually, an even weaker condition is needed; any price that is a function f of the cost of all other firms, provided that $f(a, a \ldots a) = a$ and $\partial f / \partial c_j \geq 0$ for any j, will achieve first best, where a is the cost of all other firms.

hospital- or patient-specific portion of this system as fee for service or as cost reimbursement; in this sense, a mixed system is partially a fee-for-service system and partially a fixed payment per patient.[11] This conclusion does, however, depend on the unrealistic assumption that the amount of inefficiency is invariant to the degree of cost reimbursement; to the degree that this is the case, it is clear that the weight on hospital-specific cost should be reduced.

These models implicitly assume that the hospital cannot control its admissions (i.e., they are exogenous). If this is not the case and if the hospital cannot or is not willing to cross subsidize among patients, then payment below (marginal) cost will result in the hospital's not accepting that class of patient. If hospitals can control their admissions but are willing to cross subsidize, the outcome will depend upon their willingness to cross subsidize.

At the time the PPS was established, hospitals varied substantially in the costs to the Medicare system of treating various patients. To the architects of the PPS, the bulk of this variation was attributed to inefficient hospitals that had survived because cost reimbursement had permitted it. This followed exactly the reasoning of Shleifer just described, and as already pointed out, the reformed payment system resembled "yardstick competition." The reformers recognized, however, that some of the variation in cost across hospitals was attributable to variation in the case mix of the hospital's patients; they therefore made reimbursement proportional to the case mix index, an average of the DRG weights attached to the patients at the hospital.

An ongoing concern among a number of analysts, however, has been that many DRGs have substantial within-group variance (coefficients of variation over 1, for example), and that hospitals may not be getting random draws within DRGs, as the PPS implicitly assumes.[12] Thus, the PPS may be "unfair" because hospitals whose patients are on average sicker within a DRG category receive too little reimbursement and those whose patients are on average healthier receive too much. Moreover, if the characteristics that make the patient sicker or healthier within a DRG are observable to the hospital and if the hospital can (and is willing to) control its admissions, there will be "skimming and dumping," i.e., selection effects because price does not mirror (marginal) cost at the patient level. This has led to proposals for changes in the reimbursement formula.

---

[11] There is clearly a difference between the insurer's or the government's quoting a price schedule for each service and cost reimbursement in terms of the firm's incentives to find the lowest cost suppliers of any input; that distinction, however, is not important for my immediate purposes.

[12] It is assumed that with random draws the law of large numbers will yield similar mean case mixes.

In addition to unobserved but real patient heterogeneity, hospital costs may vary for other reasons that, in principle, the government would wish to reimburse but does not because it is too costly to get the information. For example, the PPS adjusts reimbursement according to a wage index for the hospital's local area. The wage index, however, may contain measurement error; for example, the index is uniform over a metropolitan area, but inner city hospitals may have to pay higher wages than suburban hospitals (Hendricks, 1989).

Keeler (1990) terms such variation in cost "justifiable," as distinguished from cost variation attributable to variation in inefficiency, which is "unjustifiable." In Shleifer's model all cost variation is unjustifiable. Pope (1990) shows that, not surprisingly, the more the variation in hospital cost is made up of justifiable factors, the more weight should be placed on hospital-specific costs in a mixed reimbursement formula. Keeler attempted to derive an empirical estimate of the weight to be placed on hospital-specific costs in a special case and showed that it would be at least 15 to 29 percent.

The discussion here has related to hospitals, because that is the context of the PPS, but the degree of justifiable cost variation across physicians (from case mix variation) is almost certainly even greater, if for no other reason than physicians treat much smaller numbers of patients. It is probably for this reason that aggregating payment for hospitals has made more headway than aggregating payment for physicians (i.e., one rarely hears serious discussion of capitation payments at the physician level).

I now proceed to a more formal derivation of Pope's theoretical results and close this section by discussing Keeler's empirical results. Goodall (1990) makes a similar argument to Pope's and comes to the same conclusion—that payment should be a linear combination of a hospital-specific rate and an overall average rate. Those familiar with these papers may wish to omit the remainder of this section other than the last paragraph.

Pope defines a quantity, standardized average cost (SAC), which for hospital i is

$$SAC_i = T_i + e_i + s_i, \tag{3}$$

where T is treatment intensity or quality, e is an index of inefficiency, and s is an index of the unobserved severity of the hospital's patient mix or other factors that, if they could be measured accurately, would be incorporated into payment. Notice that by making T, e, and s constants, this equation implicitly assumes constant returns to scale. As Pope notes, constant returns are unlikely—a small hospital, for example, could not efficiently utilize a magnetic resonance imaging unit. To the extent that there are scale-related (or scope-related) differences in

cost that are desirable, the weight that should be placed on hospital-specific differences in costs is increased. No doubt there are such differences; the optimal size and scope of a hospital for a sparsely populated rural area would surely differ from that of a major metropolis. But Pope's desire was to show that there should be some weight placed on hospital-specific costs, so the assumption of constant returns seemed a justifiable simplification.

A strong assumption of this specification is that e_i, the efficiency parameter, does not depend upon the share of cost reimbursed; thus, hospitals derive no benefit from expenditures. A more complete analysis would have to account for any such effect.

Pope next defines a desired payment per case on the part of the government, which is

$$P^d = E(T) + s_i. \tag{4}$$

E(T) is the average cost over all hospitals, which for convenience is assumed to be the cost of efficiently producing the intensity the government wants. The important point here is that the government does not wish to reimburse for variation across hospitals in cost attributable to inefficiency; hence, e_i does not appear in this formula.

Pope now makes two key assumptions. First, the government will reimburse hospitals with a linear contract. Its payment P for each case is

$$P_i = R + \alpha SAC_i. \tag{5}$$

Note that the actual PPS, although it does contain some hospital-specific adjusters, such as the wage index in the hospital's local area, does not account for hospital-specific costs; in other words, in the actual formula, α is zero and payment is independent of SAC_i.[13] Thus, the issue here is not whether a nonlinear contract would be still better; it is whether a more general formula that allows for positive α would improve over the present formula that constrains α to be zero.

Pope's second key assumption is that the government seeks to minimize the mean square error between actual and optimal prices in its payment formula. Thus, it minimizes

[13]There is a payment based on the house staff/bed ratio and the proportion of indigent patients, but there is no direct account taken of hospital-specific costs.

$$MSE(R, \alpha) = (1/n)\ \Sigma(P_i - P^d)^2$$

$$= (1/n)\ \Sigma(R + \alpha(SAC_i) - E(T) - s_i)^2,$$

(using Equations (4) and (5)), which, rearranging terms

$$= (1/n)\ \Sigma\,(s_i - (R - E(T)) - \alpha(SAC_i))^2. \qquad (6)$$

This quadratic formula weights larger pricing errors proportionately more heavily than smaller errors, which seems reasonable, but any claims for optimality for the exact formula require a number of strong assumptions.[14]

Equation (6) is the formula for a least squares regression of s on SAC, where the intercept is $R - E(T)$ and the slope is α.[15] Using the standard least squares formula and an * to denote an estimated value,

$$\alpha^* = cov(s, SAC) / var(SAC) = (\sigma_s / \sigma_{SAC})\,\rho\,(SAC, s), \qquad (7)$$

where σ is the standard deviation and ρ is the correlation coefficient between SAC and s.

Hence, if a linear payment formula is used, the weight α on hospital-specific costs will be greater, the greater is the proportion of total variance that is attributable to unobserved severity measures and the greater the correlation between observed cost and unobserved severity. Both these results are intuitively reasonable.

More important for my present purposes, the optimal weight on hospital-specific costs is zero, as in the present system, only in two special cases:

- There is no variance across hospitals in unobserved severity, or more generally in factors that the payment scheme should recognize if it could measure them (justifiable cost differences), i.e., $\sigma_s = 0$; or
- There is no correlation between observed costs and unobserved severity or other justifiable cost differences or $\rho = 0$.

---

[14]The assumptions are (a) identical marginal valuation of services linearly declining in T; (b) constant and identical marginal cost of T; (c) E(T) socially optimal (this can be readily relaxed); (d) $T - T^*$ proportional to $P - P^d$, where T^* is optimal; i.e., treatment deviations are proportional to payment deviations; (e) social welfare is the sum of consumer surplus at the individual level less cost.

[15]Using standard notation, the least squares equation is to find the α and β that minimizes the sum of the squared residuals, or

$$\Sigma(y_i - (\alpha + \beta x_i))^2.$$

The violation of the first condition was the motivating factor for this literature in the first place; i.e., it was suspected that some of the observed variation was attributable to unobserved but justifiable factors. Put another way, Shleifer assumed identical firms, but the subsequent literature relaxed this assumption. The second condition will only be satisfied in a pathological special case; because s is a component of SAC [Equation (3)], there will be a correlation between s and SAC unless there is an exactly compensating negative covariance between s and T + e. Thus, a zero or positive covariance between s and T + e is sufficient to violate the second condition.

Of course, s_i is presumed to be unobservable; otherwise the payment system, in principle, could have taken it into account. Hence, one cannot directly estimate α.[16]

Nonetheless, Pope derived an expression for α in the special case in which T, e, and s in Equation (3) are all uncorrelated. Another, probably more intuitive way to put this special case is that if this condition is satisfied, the average "justified costs" are uncorrelated with the average "unjustified costs."[17] If T, e, and s are uncorrelated, α^*, the weight on hospital-specific costs, is the proportion of variance in SAC accounted for by s.

Subsequently, Keeler estimated Pope's expression. In particular, he regressed average adjusted cost for patients with five specific conditions on a sickness ("severity") scale[18] that incorporated information on the patient's physiologic condition not utilized in the DRG classification system. The unit of observation was the hospital. He utilized data on 6,221 patients at 296 hospitals. Accounting costs for each case were adjusted by a wage index specific to the hospital, the DRG case weight, and two other hospital-specific adjustments used by the PPS (indirect teaching and disproportionate share). Keeler then averaged both the adjusted accounting costs and the severity scale across patients (approximately 23 patients per hospital). He found that the average severity scale could explain 15 percent of the variance in the average adjusted costs. In other regressions he included additional variables that may measure justifiable differences across hospitals: Including the proportion of patients dying within five days and the

---

[16]Pope showed, however, that the reduction in mean square error from using the optimal α rather than setting $\alpha = 0$ equals the R^2 from the regression of s on SAC [i.e, the regression implied in Equation (6)].

[17]Though the condition that T, s, and e are all uncorrelated is stronger than T + s uncorrelated with e.

[18]The scale utilized from 60 to 80 disease-specific variables including APACHE II for each of five conditions (congestive heart failure, acute myocardial infarction, pneumonia, cerebrovascular accident, hip fracture) and regressed mortality within 30 days of admission on these variables. Predicted mortality was then used as a measure of sickness at admission. See Keeler, Kahn, Draper et al. (1990).

proportion with nonfatal complications raised the R^2 to 29 percent. Keeler concluded that 15 to 29 percent was a lower bound for α^* and hence a lower bound for the weight on hospital-specific costs (assuming that T, s, and e are uncorrelated). This calculation, however, assumes that payment on the basis of cost does not affect cost; to the degree it does and induces inefficiency in production, these shares are overstated.

One can think about the models with heterogeneity as a second-best problem. Shleifer showed that a lump-sum payment was optimal from the point of view of least-cost production. Relaxing the assumption of identical firms, however, raises the concern that a lump-sum payment may lead to selection. It is not difficult to construct a model in which selection is greater, the more profitable it is.[19] Thus, one trades off the amount of selection and inefficiencies in production, much as health insurance from the point of view of the consumer trades off moral hazard and risk aversion (Zeckhauser, 1970).

Ellis and McGuire and Selden: Imperfect Agency and a Mixed Reimbursement System

Whereas Pope, Goodall, and Keeler relaxed Shleifer's assumption of identical firms, Ellis and McGuire (1986) relaxed his assumption that the hospital or firm was the relevant decisionmaker and that it sought to maximize profits. In doing so, they reached a similar result to Pope's, with fully prospective payment again emerging as optimal in a special case.

In Shleifer's model there was no agency problem because the firm was a single decisionmaker. In Ellis and McGuire's model there is an agency problem. They set up a model of a physician decisionmaker who acts as an agent for two principals, the patient and the hospital, and thus has a utility function defined over patient benefit and profit to the hospital from payment for the patient. Patients are fully insured in their model and receive whatever treatment the physician orders. (They relax this last assumption in a subsequent article that I come to below.)

---

[19] An assumption of profit maximization by an insurer/provider (e.g., HMO) yields a knife-edge type prediction, whereby patients are always dumped by the HMO if their expected costs as estimated by the HMO even slightly exceed the premium or revenue. Such behavior is not observed. In order to avoid such problems, one can introduce some cost to dumping. One form of a model with such a cost is the standard supplier-induced demand model with ethics against dumping replacing ethics against inducement. In such a model as the deviation between cost and revenue rises for any individual patient, both the substitution effect and the income effect will favor dumping (unlike the supplier-induced demand model as price rises above cost), but there need not be any dumping for sufficiently small deviations (a corner solution). The more perfect market competition is, however, the more dumping will be observed.

Ellis and McGuire show the condition under which the PPS will be consistent with an efficient solution, namely that the physician is what they term a "perfect agent." Unlike some other usage, however, they do not mean by perfect agent that the physician acts as the patient would act if the patient had the same information as the physician, but rather that the physician values equally a dollar's worth of benefit to the patient and a dollar's worth of benefit to the hospital.[20] If this is not the case—specifically if the physician values a dollar's worth of benefit to the hospital more than a dollar's worth of benefit to the patient—then it is optimal to have reimbursement be a linear combination of a cost-based system and a prospective system, with the weight on the prospective system equal to the value the physician places (at the margin) on a dollar of benefit to the hospital relative to a dollar of benefit to the patient. They call this "supply-side cost sharing," indicating that the hospital is at some risk of bearing excess treatment cost through the prospective payment.

Intuitively, if the physician trades off dollar benefits to the hospital and patient equally, one reaches an efficient solution with a lump sum per case because the physician internalizes social benefits (as defined in this model) properly, whereas if the physician overweights hospital profits one can compensate for the physician's preferences to underdeliver services by reducing the hospital's profit rate (i.e., not have the hospital keep 100 percent of any savings from the underdelivery of services).

I now proceed to show this result more formally; this repeats Ellis and McGuire's formulation, so readers familiar with their paper can omit the remainder of this subsection, except for the remarks about their model that close the section. In the Ellis-McGuire model the patient is fully insured and accepts the recommendation of the physician. The patient has a (total) benefit function $B(q)$, where q is quantity of hospital services. The marginal benefit function is $b(q) = B'(q)$. $b(q)$ can be negative from medical error. $B(q)$ measures the full social benefit; there are no externalities, though this is not essential to the conclusion that a mixed reimbursement system will in general be better than a pure system. By definition the profit to hospital $= \pi(q) = R(q) - C(q)$.

Physicians, who make the resource allocation decisions and, in particular, decide on q, all have utility functions $U(\pi(q), B(q))$.[21] The first order condition for physicians is

---

[20]This is analogous to maximizing the sum of consumers' and producers' surplus, but benefits to both patient and hospital are mediated through the physician's utility function.

[21]Note the analog with Pauly and Redisch (1973).

$$U_\pi \pi_q + U_B B_q = 0. \tag{8}$$

Define

$$MRS_{\pi,q} = U_B b / U_\pi = (MRS_{\pi,B})(b) = \alpha b, \tag{9}$$

since $b = B_q$ and defining $\alpha = MRS_{\pi,B}$. α thus equals the rate at which the physician is willing to trade off a dollar of benefit (profit) to the hospital for a dollar of benefit for the patient or the degree to which the physician takes the patient's interests into account.

Under cost-based reimbursement revenues equal cost by the definition of cost-based reimbursement, so $\pi = R(q) - C(q) \equiv 0$; hence, $d\pi / dq = 0$. Since $d\pi / dq$ always equals zero, a physician treating a fully insured patient will maximize the utility function by always behaving so as to make $dB / dq = 0$; i.e., the physician will carry out all procedures with expected positive benefit irrespective of cost. Thus, there are too many resources in hospital production relative to the socially efficient amount.

Turn now to prospective payment. Assume first full prospective payment at the patient level with revenue fixed at a; hence $R(q) = a$.

$$\pi = a - C(q). \tag{10}$$

Differentiating Equation (10) with respect to q and letting marginal cost be c(q), we have using Equations (8) and (9)

$$\alpha b(q) = c(q) \tag{11}$$

[since from Equation (8), $\partial\pi / \partial q = -U_B b / U_\pi$ = [using Equation (9)] $- \alpha b$].

In the simplest case, where marginal cost is constant, we can write $\alpha b = c$.

Suppose $\alpha = 1$ (perfect agency). Then from Equation (11), in the case of constant marginal cost, $b = c$, or the marginal benefit to patient equals the marginal cost to hospital. If payer sets payments prospectively so as to cover costs from the cost-based system, as Medicare did, then hospitals will initially make profits, while they cut back intensity, but Medicare can recoup those profits by lowering rates (assuming it knows about the profits and hospitals do not bury them in their cost structure). Ellis and McGuire characterize this result as "the promise of prospective payment." Put another way, under these conditions the socially efficient quantity of resources is devoted to hospital services, unlike cost-based reimbursement.

Suppose now that agency is not perfect. If $\alpha < 1$, there is too much weight given to hospital profits relative to patient benefits. ($\alpha = U_B / U_\pi$ so if $\alpha < 1$ an additional dollar of benefit to the patient generates less utility for the doctor than an additional dollar of profit to the hospital.) Then using Equation (11), there will be too little benefit to the patient since b will be greater than c, or, at the stopping point, the marginal benefit to the patient will exceed the marginal cost. This is the concern about undersupply in a fully prospective system.

Ellis and McGuire then turn to a two-part reimbursement system (a linear contract) in which:

$$R(q) = a + rC(q), \tag{12}$$

where a is a fixed payment as in Equation (10) and r is a fraction between 0 and 1 that represents the fraction of costs reimbursed. In this system, the fixed amount a can be lowered as r is increased in order to maintain the same total payment per case (budget neutrality). Ellis and McGuire then show that this mixed system can induce the socially efficient level of supply when the physician is an imperfect agent as follows:

By definition profit $\pi = R(q) - C(q)$ and the physician's problem is to maximize [using Equation (12)]:

$$U(\pi, B) = U(a + rC - C, B) = U(a + (r-1)C,\ B).$$

Assume a linear cost function (constant marginal cost) and constant marginal benefit, differentiate with respect to q, and set the result equal to 0:

$$U_\pi(r-1)C_q + U_B B_q = U_\pi(r-1)c + U_B b = 0.$$

Since $\alpha = U_B / U_\pi$, we have

$$(r-1)c = -\alpha b,$$

or

$$(1-r)c = \alpha b. \tag{13}$$

For efficiency we require that b = c (marginal benefit equals marginal cost), so the optimal r is defined by

$$(1-r) = \alpha, \text{ or } r = 1-\alpha \tag{14}$$

$(1-r)$ is the cost sharing on the supply side, i.e., it is the portion of the cost borne by the hospital, so the optimal cost sharing goes down as the degree of agency goes down. The assumption of perfect agency ($\alpha = 1$), which yields full prospective payment as the solution, is a special case of this formula. With

perfect agency, the physician will weight a dollar of benefits to the patients equally with a dollar of benefits to the hospital, so it is appropriate that the hospital bears the full marginal cost. As the physician becomes more of an agent for the hospital, this is offset by decreasing the cost sharing to the hospital.

How would one implement these notions? Specifically, how might one estimate α and hence define r? Suppose c is a constant. For a given supplier, α can be regarded as constant. Denote incremental payment to the hospital per case, rc, as p and differentiate Equation (13) with respect to p.[22] At $b = b(q)$,

$$-1 = \alpha(db/dq)(dq/dp), \text{ or dividing through by } -\alpha(db/dq)$$

$$dq/dp = -1/(\alpha(db/dq)) = -(dq/db)/\alpha. \qquad (15)$$

The left-hand side can be thought of as a sort of supply curve (how the amount delivered changes with incremental payment). This equation states that the slope of this supply curve equals the negative of the slope of the demand curve (since b is the marginal benefit or the price the consumer is willing to pay) divided by α. Multiplying both sides by p/q converts the equation into elasticities. If $\alpha = 1$, this says that the elasticity of supply will equal the negative of the elasticity of demand. Hence, if the marginal benefit curve equals a demand curve and the cost curve is linear, the magnitude of α can be inferred by comparing the supply response to prospective payment with what we know about demand elasticities.

Ellis and McGuire (1990) show that analogous results arise in a different model, which relaxes the assumption of a passive patient. In this new model both patients and providers have preferences over the amount of care, although the provider continues to care about both hospital profits and patient benefits. In general, patients' preferred amounts of care differ from providers' preferred amounts, and there is a bargaining solution that is a Nash equilibrium.[23] In the special case that they consider, the optimal payment is again a linear contract, although only in a special case is it the same linear contract as in Ellis and McGuire (1986).

Three remarks might be made about these results:

1. Ellis and McGuire assume a linear contract; a nonlinear contract might be still better, but this remains to be shown. Nonetheless, the linear contract

---

[22] $(1 - r)c =$ with new notation $(c - p) = \alpha b(q)$.

[23] Ellis and McGuire illustrate their solution for a special case in which the parties have equal bargaining power.

that Ellis and McGuire consider is certainly more general than a pure fee for service or a pure prospective system.

2. Ellis and McGuire show that, not surprisingly, with risk aversion on the part of the patient but not the provider and an actuarially fair insurance policy with a constant coinsurance rate, the optimal coinsurance rate for the consumer is zero. In other words, the linear contract, which they refer to as supply-side cost sharing, is preferable to demand-side cost sharing of the traditional type. This result, however, requires a number of assumptions, some of which seem relatively weak, but others much less so:

   a. That the consumer's insurance policy takes the form of a constant coinsurance rate and the premium for this policy is actuarially fair. It should not be surprising that the availability of actuarially fair insurance to a risk-averse consumer yields a result that full insurance for the consumer is optimal. If, however, the policy can be in the form of a deductible with coinsurance (perhaps zero) above the deductible and insurance carries a positive loading, surely the more relevant case, it seems likely that some demand-side cost sharing would be optimal.

   b. That α is ≤ 1; if α is > 1, demand-side cost sharing is necessary to "restrain" the provider. At an extreme, a physician who considers only patient preferences would have $\alpha = \infty$. This may seem implausible (though it is what some literature describes as perfect agency), but if there is competition among physicians for patients, such competition may force physicians to take account of patient preferences sufficiently that α might be greater than 1.

   c. That the provider is not risk averse. This is plausible for a rather large hospital, but for a single physician or even for a small hospital it is less plausible, at least given the heterogeneity within existing case mix categories.

3. The assumption that the physician is the only agent does not seem to capture the reality that hospital and physician incentives may differ. For example, the payment under the PPS may be such that the hospital would lose by taking a very sick patient, but the physician would gain (because the physician could perform many procedures that would be reimbursed above cost under fee for service). The Ellis and McGuire model is in the spirit of the Pauly and Redisch (1973) paper of the hospital as a physician's cooperative, but there are other possible claimants for a hospital surplus. Thus, it may be useful to generalize this model to consider the hospital, the physician, and in some cases the patient as independent decisionmakers.

Selden (1990) derives an analogous result to Ellis and McGuire in the context of capitation (a health maintenance organization) rather than in the context of the PPS.

Newhouse: Pricing Errors and a Mixed Reimbursement System

My own 1991 paper relaxed Shleifer's assumption of competitive pricing and yet another rationale for a mixed reimbursement system emerged. For reasons explained in the introduction, I argued that competitive pricing by insurers was improbable. Based on that observation, I developed an argument for a mixed basis of pricing in the context of capitation and fee for service, but the argument could apply equally to the PPS context that we are considering.

The argument is that in reality we are dealing with administered price systems that set prices to approximate some ideal price but do so only with error. Prices can be set for various bases (e.g., capitation, price for a DRG, or for each service). Provided the errors made by the insurer in these different bases are not perfectly correlated, there will be a gain from averaging the errors; i.e., from using multiple bases of reimbursement.

I now make these ideas somewhat more formal. For simplicity suppose that demand and supply curves are linear.[24] Let there be a set of demand and supply curves corresponding to a PPS (or capitation) world and another corresponding to a fee-for-service world. There is a set of prices an insurer sets corresponding to each world.[25] For each basis of pricing there is an optimal price that the insurer attempts to approximate.

Random but unbiased errors in the insurer's prices create a welfare loss that is measured by the usual triangle, the size of which increases as the square of the error.[26] Hence, averaging the errors will minimize welfare loss for the same reason that averaging additional observations reduces the variance of a mean. In

[24]The algebra is simple with linear supply and demand curves, but all I need to show that a pure system is not optimal is that there exists an optimal price in each base (cost based, fee for service, and prospective) that the insurer attempts to set but makes (not perfectly correlated) errors. For example, suppose the population is informed but that there are externalities, and let the demand curve be the marginal valuation curve across the entire population. Suppose the insurer is trying to obtain efficient production; then the insurer wishes to set the supply price where that valuation curve intersects the marginal cost curve, but the insurer may make errors in doing so.

[25]In the United States the set of prices corresponding to a capitation world could be taken to be Medicare's Adjusted Average per Capita Cost, and the set of prices corresponding to the fee-for-service world could be taken to be the PPS for hospitals and the Resource Based Relative Value Scale for physicians.

[26]One can introduce bias without changing the nature of the result. See Newhouse (1991).

the simple case of linear supply and demand curves and unbiased errors, the weights on the various bases of payment should be inversely proportional to the variance of the error in that base. For example, if capitation errors have the same variance as fee-for-service errors, the price should be a simple average of payment under each system. Intuitively, one downweights the basis of pricing that is more susceptible to error. Similarly, if setting the price paid per admission in the PPS has the same error variance as prices paid under a more disaggregated fee-for-service system, the two payment systems would be averaged.

Siegel et al.: Estimation Theory and a Mixed Reimbursement System

Siegel et al. (1992) also advocate a mixed reimbursement system, one blending a national average rate, a hospital-specific rate that could reflect both the efficiency of the hospital and unobserved case mix factors, and observed patient-specific factors, such as age. The basis for their recommendation, though analogous to Pope, Goodall, and Newhouse in its use of squared error loss as a criterion, is grounded in estimation theory and the assumption (not necessarily consistent with the models just discussed) that a blended rate will minimize inefficiency (that e_i in Pope's notation will be small).[27] On this assumption, the hospital should be paid its cost for treating each patient, which using Pope's notation equals $T_i + s_{ij}$, where i indexes the hospital and j the patient.

Siegel and her colleagues draw a link between James-Stein estimation and the problem we are considering. James and Stein showed that for an ensemble of three or more elements, the sample mean is not admissible (i.e., does not minimize mean square error). To illustrate, take a variant of our problem and suppose that one wants to estimate average cost at four hospitals. The James-Stein result says that one can do better than simply using the mean cost of each hospital as an estimator. In particular, certain linear combinations of the mean of the ensemble (e.g., the national mean of all hospitals) and the mean of each element of the ensemble (e.g., the individual hospital), or so-called shrinkage estimators, have less mean square error for the ensemble (the estimates of the four hospital means) than the sample mean for each element (Efron and Morris, 1973, 1975). Under certain assumptions, a better estimator is $\lambda_i\mu + (1 - \lambda_i)C_i$, where μ is the national mean, C_i is a hospital-specific mean, and

[27] Siegel et al. suggest that over time e will converge toward zero because the government will learn what costs are. As noted in the discussion of Laffont below, this is a doubtful assumption because firms may anticipate the government's future actions. Nonetheless, the estimation theory that Siegel et al. bring to bear suggests that there may be a gain in using a linear contract with some positive weight on hospital-specific costs if one is trying to estimate provider-specific means.

$\lambda_i = \sigma_i^2/(\sigma_i^2 + r^2)$, where σ_i^2 is the variance of the hospital mean, C_i, and r^2 is the variance across hospital means (i.e., the between-hospital variance). Thus, as a sample at a hospital becomes large, the weight on that hospital's mean tends toward one (λ_i tends toward zero), and as between-hospital variance tends to zero (all hospitals alike), the weight on the grand mean tends to one.

Siegel et al. assume that the government wants to pay $X_{ij} = T_i + s_{ij}$, which is a patient-specific cost. In this sense they assume only "justifiable" cost differences among hospitals because the government will learn about and eliminate unjustifiable cost differences. They show that this leads to the types of linear contracts we have been discussing; that is, even if one were attempting to pay a patient-specific mean, one would use a mixed reimbursement system that was not the simple hospital average. For simplicity, consider patients in one DRG.

Siegel and her colleagues actually consider a slightly more general reimbursement formula of the form: $C_j + a_jX_{ij} + b_j\mu$, where C_j is the average cost for the DRG at hospital j, X_{ij} are observed treatment costs of patient i at hospital j, and μ is a national mean cost for the DRG. They show that if one wants to minimize expected quadratic loss for matching cost and revenue across all patients,

$$E[X_{ij} - (C_j + a_jX_{ij} + b_j\mu)]^2 = E[C_j - ((1 - a_j)X_{ij} - b_j\mu)]^2, \tag{16}$$

one should choose the following payment formula:

$$C_j + \lambda_j(X_{ij} - \mu), \tag{17}$$

where λ_j is the ratio of the within-hospital variance of the X_{ij} to the sum of that variance plus the between-hospital variance of the C_j. Hence, in Equation (16), $a_j = (1 - \lambda_j)$ and $b_j = -\lambda_j$.[28] Note that this result is in a sense a dual of the usual James-Stein result, since the shrinkage factor on the X_{ij} is $(1 - \lambda)$ rather than λ. In the case of $\lambda = 0$, meaning no within-hospital variance in patient characteristics, this is cost reimbursement; if $\lambda = 1$, meaning no between-hospital variance, this is cost reimbursement plus the difference between the (scaled) patient characteristics and the national average.

Siegel et al. illustrated their method with an application to 60,000 psychotic patients being treated in 800 facilities. They estimated λ_j and used the patient-specific factors of age, gender, race, and disability status to predict cost X_{ij}. The

[28]Let the usual shrinkage estimator of the X_{ij} be $\lambda_j\mu + (1 - \lambda_j)X_{ij}$ and suppose we seek to minimize the expectation of the squared difference between this estimator and C_j. One can readily verify that this difference equals $X_{ij} - [C_j + \lambda_j(X_{ij} - \mu)]$, where the term in brackets is the expression in Equation (17). Hence, payment is set equal to the term in brackets.

λ_j varied across facilities from 0.003 (indicating almost no within-facility variance in patient-specific predicted cost) to 0.97, with a median of 0.58.

In a sense, Siegel et al. could be viewed arriving at a mixed system by modifying cost reimbursement at the patient level rather than by modifying full prospective payment at the hospital level. That is, they in effect solve the following problem: If one wanted to pay the expected cost of each patient rather than the average cost of all patients (perhaps to avoid selection against high-cost patients or to reimburse hospitals more who had more high-cost patients) and one observed some characteristics of patients that predicted cost (e.g., age, diagnosis, etc.), one would be better off in the sense of mean square error not to pay simply the predicted cost for each individual patient but rather to pay a linear combination of the predicted cost for each patient, the mean cost of all patients in that hospital, and the national mean. In general, the theory of James-Stein estimation suggests a mixed rather than a pure reimbursement system.

Laffont: The Procurement Problem with Nonobservable Effort, Nonhomogeneous Firms, and Bidding

I now turn back to the industrial organization literature and show that a generalization of Shleifer's model in that literature by Laffont (1987) also reaches the conclusion that a mixed reimbursement system is preferred to a pure reimbursement system when there is bidding, as is the case in selective contracting states. (Similar results to Laffont, 1987, are obtained in Laffont and Tirole, 1986.) In this model there is hidden information from the government about unobservable factors that influence a hospital's costs (e.g., case mix), including its effort in minimizing cost. Thus, as in Shleifer's model, firms can reduce costs by making additional effort, but those efforts are not observable by the government/regulator/insurer (e.g., costs can be reduced by managers' working harder but their work effort is not observed). The hospital, however, reveals something about its cost by submitting a bid to care for patients.

The Laffont model generalizes Shleifer's model in three ways:

- Rather than being identical, hospitals are heterogeneous in ways the regulator does not fully know; additionally, there is a random element in observed (after-the-fact) cost. The government desires to pay these cost differences.
- Unlike Shleifer's model, taxes to finance payments to the hospital or firm are not lump sum but involve a loss of efficiency; and

- Unlike Shleifer's model, there is a bid submitted by the hospital. Bids matter for the government's payment to a hospital as follows: The government will pay a hospital's announced cost or bid plus it will share in deviations between the bid and the final observed cost (overruns and underruns). The government's share of overruns and underruns can vary according to the level of the bid. The hospital, knowing the government's reimbursement rules, announces a cost it expects it will incur. This announced cost or bid, however, is restricted to lie in a certain interval because the government is assumed to know that cost cannot exceed a certain amount nor be less than a certain amount.

  Both the hospital and the government know that the higher the announced cost, the more the government will share in overruns and underruns. For a hospital that announces cost at the lowest end of the interval into which costs can fall, the government will use a fixed-price contract, because the government believes costs cannot go lower and with a fixed-price contract the hospital bears any overrun. Conversely, for a hospital that announces a cost at the highest part of the interval, the government will use a cost-reimbursement contract, because costs cannot go any higher and the government will reap any savings if they are lower.

Thus, this model leads to a contract in the form of Equation (17); as observed cost, C, rises, the payment to the hospital increases.

A hospital or firm has two arguments in its utility function: (1) It can gain in the usual sense of profit by reducing its actual costs, and it will gain more, the greater is the sharing fraction. Because the hospital's share of any underrun increases as its bid falls, this is an incentive to bid low; (2) It gains if it has to make less effort at cost reduction; given a sharing fraction, to achieve the same profit, it will have to make less effort the higher the bid; this is an incentive to bid high.

In Laffont's model, then, hospitals have different true cost parameters; in addition, hospitals may make varying efforts to keep costs down, and there is furthermore random variation in observed cost. Both because there is true cost variation and because there is a random element to cost, the payer (the government in Laffont's model) cannot infer the hospital's effort from observed cost.

Laffont makes the simplifying assumption that the government is trying to purchase a good (in our case hospital care) whose benefits exceed the costs of production no matter which hospital produces it (subject to a constraint that cost cannot exceed a certain figure); hence, the hospital should always be in business

and the only issue is the price the government will pay the hospital.[29] The quantity dimension is effectively kept outside the model by assuming that only one unit of the good will be produced; by rescaling, this is equivalent to assuming that quantity is exogenous and that there are constant returns to scale.

One might ask why this model applies to our problem of how a government or insurer should reimburse hospitals (or other health care providers); i.e., why would not the government simply seek the cheapest hospital? One answer to that question is that patients may have preferences for specific hospitals, perhaps for religious reasons or because they live nearer them, but another more important reason is that costs may vary for justifiable reasons as described above, that is, for reasons that the government would be willing to reimburse.[30] These variations in cost across hospitals may be random in any one year (e.g., the heart attacks that came to the hospital this year were unusually severe or not severe) or may be constant from year to year (e.g., the hospital installed a piece of equipment with which it treats all those with a heart attack and the government believes this particular hospital should have that piece of equipment, whereas another hospital should not because it could not use it at an efficient scale). I come, however, below to a variant where the government seeks bids and awards the contract to the lowest bidder.

Effort to reduce cost or inefficiency is valued negatively by the firm. Therefore the firm must be compensated for the effort, but the regulator cannot observe the firm's effort. (In Shleifer's notation, R is not observable.) This would not matter if firms were identical—or differed only in observable ways; e.g., if all interhospital variation in case mix were captured by the case mix index—the government would only have to ensure that its payment included R.[31] For that purpose one would only have to set R at a level that kept the firms in business, which one might approximate by trial and error. Thus, even if R is not observable, a pure PPS (i.e., not a mixed system) is optimal if firms are homogeneous (this is Shleifer's result).

Laffont's model also differs from those considered up to now in that he explicitly considers the distortions that arise in raising government monies to procure

---

[29] In Laffont's problem the government is trying to have the firm announce costs that are truthful. Higher cost firms will have more of an incentive to do so if the government shares those higher costs, which is the effect Laffont's scheme has.

[30] The government might ask consumers who wish to use higher priced hospitals to pay the difference, but they are presumed not to want to charge more to consumers who use hospitals that treat sicker patients.

[31] The condition on all variation being observable can be relaxed to be that all remaining variation is random with a mean of zero (or an observable mean that can be reimbursed) provided the hospital (firm) is risk neutral.

goods or services. Thus, his model clearly applies to tax-financed medical insurance. Employer-financed medical insurance creates other distortions, for example, possible job loss around the minimum wage; hence, Laffont's model adapts straightforwardly to employment-based insurance. In Laffont's model efficiency requires accounting for the distortionary costs of financing.[32]

This brings us to the crux of the problem. Laffont's model is about structuring a contract that will minimize the total social cost of producing the good or service; i.e., will trade off the loss from inefficient production caused by too little effort in cost reduction, which stems from not using a fixed-price contract for each hospital (firm), against the loss from not minimizing rents, thereby having taxes higher than necessary. Laffont shows that the government can procure the product at a lower overall cost to it, thereby minimizing taxes or premiums, if it shares costs with firms. Doing so, however, leaves firms with an incentive not to carry cost-minimizing effort to the point that would be socially desirable. Laffont's model suggests that in general the government should share costs and the sharing fraction should depend upon—in our problem—observed cost. I now make Laffont's model more formal.

Let a hospital's or firm's costs of production for one unit of production C_i be given by:

$$C_i = \beta_i - e_i + \epsilon_i, \tag{18}$$

where β_i is a firm-specific parameter of cost, e_i is a term reflecting the effort a firm puts into reducing cost (scaled so that a unit of effort reduces cost one unit), and ϵ_i is a stochastic term with a mean of zero.[33] The firm knows β_i, but the regulator does not. The regulator does know, however, that β_i lies in an interval bounded by β_L and β_U.[34]

In putting forth effort to reduce cost, the manager or owner of the firm incurs disutility represented by a function $D_i(e)$, which has positive first and second derivatives. Thus, the social cost of production is $C_i + D_i$. As with Shleifer, distributional considerations are ignored. A necessary condition for efficiency is that social cost be minimized. Using Equation (18), differentiating $C_i + D_i$, with respect to e_i, and setting the result equal to zero yields the first best condition:

---

[32]Note that in the previous models there was no such feature. Shleifer explicitly assumed lump-sum taxation and distributional neutrality; hence, a dollar paid from the "insurer" to the "hospital" was simply a transfer and did not change social welfare. Although Laffont also assumes distributional neutrality, he does not assume lump-sum taxation.

[33]This cost function is similar to Pope's.

[34]Laffont's derivations assume that β has a uniform distribution over this interval.

$$D_i = 1. \tag{19}$$

This is analogous to the condition in Equation (2), but in the case Shleifer considered, all firms were identical, and there was no random term; hence, $C_i = C_j$ and $D_i = D_j$ for all i and j, and the first order condition in Equation (19) yields the same effort for all firms, and all firms have the same cost.

Laffont, however, considers the case in which costs and disutilities differ across firms and in which the regulator cannot take observed cost as an indicator of effort because β and ϵ vary across firms. For simplicity in what follows I ignore ϵ_i.

The regulator could guarantee an efficient amount of effort (efficient production) by utilizing a fixed-price contract. To keep all firms in business, however, the fixed-price contract would have to pay each firm β_U. In that case, however, the regulator would pay all firms other than the highest cost firm a rent equal to $\beta_U - \beta_i$.[35] Because rents must be financed by taxes and because there is a loss of efficiency on each dollar raised through taxes of λ, the rent paid each firm causes a loss equal to $\lambda(\beta_U - \beta_i)$. Estimates of λ from the literature vary, but a middle-of-the-road estimate of λ might be 0.3 or less (Ballard and Fullerton 1992; Ballard, Shoven, and Whalley, 1985; Browning 1987; Fullerton 1991; Stuart 1984).

Under these assumptions the optimal second best reimbursement scheme is a linear contract, which comes from the solution to:

$$D'(e) = 1 - [(\lambda/(1+\lambda))\,(\beta - \beta_L)\,(D''(e))] = 1 - M, \tag{20}$$

where M is the entire set of terms in brackets. At an optimum for the firm, D'' is positive (the firm wishes to minimize D); hence, M is positive (except for the lowest cost firm, in which case M is 0 since in that case $\beta = \beta_L$). Because D' will equal 1 under a fixed-price contract, but will be less than that if the government shares overruns and underruns, the optimal contract is not a fixed-price contract (except for the lowest cost firm) and the optimal effort is less than under a fixed-price contract. Firms with costs below β_U earn rents that rise as their costs fall below β_U. Specifically, the rent of a firm with cost parameter β_i is:

$$\int D'[e(\beta)]\,d\beta, \tag{21}$$

[35] Plus the firm earns a surplus on any effort used to reduce costs below β_i.

where the integral is evaluated between β_i and β_U. Thus, the highest cost firm earns no rent because the government uses a full cost-reimbursement contract for it.

It follows from Equation (20) that the optimal sharing parameter (weight on firm-specific cost) is a function of λ; the greater λ, the greater is the weight on the firm-specific cost. In the context of our problem this means inefficient hospitals or physicians who deliver "too many" services have a *greater* share of their reimbursement based on the cost of services they deliver; less of their payment is prospective. Indeed, only the most efficient hospital or firm receives reimbursement based solely on a lump sum (capitation).

This may seem counterintuitive, but the economic intuition is as follows: The regulator can guarantee production by all firms only by setting a fixed-price contract at β_U. Doing so, however, would give all firms, except the highest cost firm, rents, and the greater λ ; is, the greater the social cost of those rents since they must be financed through taxes. Hence, the greater λ is, the less the regulator is inclined to give rents, which means accepting less effort and more inefficiency in production. Similarly, the greater β_i is, the greater the government's share of variation in cost. As already noted, if a firm announces a cost equal to β_L, the government will want to use a fixed-price contract; the more the firm's announced costs rise toward β_U, the more the government will want to use a cost-reimbursement contract.

Laffont points out that as the government's information about the range of justifiable cost $[\beta_U - \beta_L]$ improves (i.e., as the range narrows), social welfare will improve because the government will have to pay less rent.[36] One might think about the government's information improving in two ways. First, one might relax the assumption that all firms must obtain a contract; in that case, one would contemplate competition as a way of eliciting information about the firm's costs. In equilibrium competition will elicit a bid from the best firm at the cost level of the second best firm.[37] As the number of competing firms grows large, the cost will converge to the lowest cost firm and the sharing will converge to zero (i.e., the government will use a fixed-price contract), as in Shleifer's model. One can also relax Laffont's assumption that the government wants all firms to remain in business. The government or insurer does not have to pay the costs of the higher cost firms if it is willing to see those firms go out of business; thus β_U could

---

[36]In the special case in which $\beta_U = \beta_L$, we have homogeneous firms, and the government does not have to pay any rent; this yields Shleifer's result.

[37]This result assumes risk neutrality.

represent the costs of the highest cost firm that the government was willing to keep in business.

Indeed, one could reasonably ask why the government (or insurer) should try to keep higher cost hospitals in business; should one not want to see low-cost hospitals expanding their market share and high-cost hospitals leaving the market? Efficiency might well be improved by having some higher cost hospitals go out of business; however, there are some factors suggesting that even at a (practical) optimum, there would be variation across hospitals in cost:

1. Most important, costs may vary with scale, and optimal scale may vary with market area. It is not likely optimal to have all patients in the country go to one size hospital.
2. Related to this point, consider two adjacent counties, each served by a single hospital, with one hospital more efficient than another. If one accounted for travel costs of patients and friends, it may be efficient to keep both hospitals operating.
3. Even within local market areas, patients may have preferences for particular hospitals, for example because of religion or because their preferred physician has admitting privileges there. Thus, patients may be willing to pay some additional costs to use a more expensive hospital.
4. The quality of the hospital product (either the technical quality or amenities or both) may vary, and an efficient solution might allow for some variation. This is more likely, the less important externalities are (your willingness to pay for my hospital care).

One might also consider the play of Laffont's game over time. Suppose the government contracts with a hospital in year 1 and observes costs. Now suppose that the government, in announcing terms for year 2, takes advantage of the information it has on observed cost in year 1. Laffont shows that unless the government can firmly commit to not exploiting the information it gains in year 1 for its purposes, which in practice would seem difficult, there may not be an equilibrium. Note, however, that the PPS does effectively commit the government; a hospital's observed costs in year t have a negligible effect on its reimbursement in year t + 1.

To sum up, Laffont's model gives additional economic content to the arguments of Pope, Goodall, and Keeler described above. In those models hospitals may have costs that differ for "justifiable" and "unjustifiable" reasons. In Laffont's model this is analogous to firms with varying β_i (justifiable) and e_i (unjustifiable), where the government observes only the sum. In both cases, one reaches the

conclusion that in general the government should use a mixed contract to reimburse the hospital.

Lazear and Rosen: Reimbursement as a Tournament

In addition to the results from the industrial organization literature, there are results from the labor economics literature that are related to our problem. Lazear and Rosen (1981) showed that pay of workers should be based on *relative* performance if three conditions are satisfied: (1) the effort of workers is not observable; (2) the relationship between input and output is stochastic; and (3) the stochastic elements that affect the behavior of different workers are correlated.

This maps directly to Laffont's model [see Equation (18)] with the additional wrinkle that the stochastic term ϵ_i for the i^{th} hospital is correlated with the stochastic term for the j^{th} hospital ϵ_j. Under these circumstances, by looking at the performance of one worker (hospital) relative to another, one can make better inferences about effort. See also Nalebuff and Stiglitz (1983).

The economic intuition is that if worker effort can be costlessly measured, payment based on input (e.g., time) is efficient, but if effort can only be measured at high cost, one invites shirking. In the hospital context, if one can measure the effort of hospital managers to be efficient, one can base reimbursement on that effort, but in fact measuring effort directly is very difficult and simply reimbursing cost invites less than optimal effort at cost reduction. In the worker context one may be able to improve matters by relating compensation to output (e.g., piece rates). However, if output is stochastic, this imposes risk on the worker.[38] Paying by relative rank may lower the monitoring costs (because only the ordinal position must be known) and so may improve matters.

Lazear and Rosen show that payment by relative rank can induce first best outcomes. How might one apply their model to the medical reimbursement case? Instead of the type of linear contracts we have been considering to this point, their model would suggest instead rewards or penalties based on relative rank, say, among a group of hospitals. Thus, those hospitals judged superior (e.g., low-cost) hospitals might be rewarded not with a fixed-price contract but

[38] In Stiglitz (1974) this trade-off between risk and shirking led to a linear contract of the type we have been discussing between a landlord and a farmer; however, it is not clear that risk should play such a prominent role in the case of hospital reimbursement. Risk for individual physicians, however, would be higher than for hospitals, so this literature appears more relevant to physician reimbursement.

with a prize, or high-cost hospitals might be penalized; unlike the models we have been considering to this point, however, the prize or penalty is not related to the amount by which the hospital differed from the mean. In other words, rank, not absolute differences, matters.

The potential gain from "grading on a curve" would occur if one believes (as I do) that the measured accounting costs are only imperfectly correlated with true economic costs.[39] This imperfect correlation removes one rationale for relating payment as tightly to observed cost as a linear contract implies, while comparing the hospital to the performance of others allows for some incentives not to shirk. (Recall in Pope's model that as ρ falls, the weight on hospital-specific costs falls.) The economics of this in the medical context, however, remain to be worked out.

Conclusion

My main point is that the debate over what is the "best" way to pay the doctor or hospital usually presumes that a pure system is the "best" way. In fact, there are many reasons for believing that one can do better with a mixed system rather than a pure system. At an abstract level, a pure system is a special case of a mixed system, so ignoring the trouble to administer the system, one should always be able to do at least as well with a mixed system. The more practical issue is whether it is worth any additional cost of setting up a mixed system. I believe it is, in part because the additional costs for hospitals is small; the information is already collected.

The argument can be put another way. The proof that a pure system, such as the PPS, is optimal requires some unrealistic assumptions, most important, that all providers (hospitals) are homogeneous or that any "justified" variation in cost among them can be accounted for by observable variables, such as the Case Mix Index (an average of DRG weights).[40] Because we know there is substantial within-DRG variance in patient cost that is related to unobserved case mix factors, the necessary assumption does not seem plausible, and relaxing this assumption leads to reimbursement that takes some account of either the hospital's observed cost or the actual services delivered to the patient. Either way, reimbursement is not entirely prospective. Existing outlier payments are a form of accounting for actual cost, but I believe too modest a step.

---

[39]For example, the accounting costs usually measure capital costs very poorly; they also are average costs; however, one wants marginal costs.

[40]Note also that the cost-reimbursement system, which PPS replaced, was also a pure system.

In the context of a capitated system, one can make the same arguments; either reimbursement is partially a function of the organization's observed costs or the actual services it delivers to its patients. Much has been written about "risk adjusters," such as age, in a capitated system. In the view here, one risk adjuster would be the volume of services delivered.

Two critical problems remain:

1. The optimal—or even sensible—weights on the various bases of payment must be determined. Although the papers discussed frequently derived theoretically optimal weights, the derivations required special assumptions and in addition were often based on unobservable quantities. To recapitulate:
   - Pope, Goodall, and Keeler use as a weight the proportion of the residual variance in cost that is justifiable; this will generally not be observed.
   - Ellis and McGuire require the agent's (physician's) marginal valuations of benefits to the hospital and to the patient.
   - My model requires the relative error variance of pricing errors.
   - Siegel et al. assume that inefficient production if providers are reimbursed at cost is not a serious problem, though many of the other models (Shleifer, Laffont) are about precisely that problem. Under Siegel et al.'s assumption, and with the additional assumption of normality, one requires only within and between variances, which in principle are observable.
   - Laffont requires knowing the true distribution of cost, as well as numerous other special assumptions.

Thus, it remains to be shown that in real-life contexts one can actually derive weights that improve matters. (And, of course, one has the problem of determining fees or capitated rates that are optimal.)

Certainly not all mixed systems will improve over a pure system. For example, suppose it were known that the optimal weight on hospital-specific costs were 0.01. In that case it seems likely that the current U.S. PPS, which gives hospital-specific costs no weight at all, would be better than a mixed system that gave them a weight of 0.99. In the absence of an implementable method that is guaranteed to improve matters, we must fall back on informal trial and error;

that is, move incrementally away from a pure system and observe the result. If the result is an improvement, move an increment further away.[41]

2. The basis of payment must be determined. Probably the most aggregate basis of payment is a pure capitation. At or near the other extreme is the current American system for paying physicians, which defines several thousand services, each of which is priced separately. In this report I have treated the payment problem either as one of compromising between these two bases of payment—or in the hospital context of paying a lump sum for the entire admission versus a fee for each service during the admission. But as the PPS demonstrates, in between the extremes of capitation and fee for service are varying levels of aggregation such as a fixed payment for an entire stay in a hospital as in PPS. Likewise one might pay for an entire illness episode, for example, a lump sum for an operation including preoperative and postoperative care or a lump sum per year for certain chronic diagnoses or combinations of them. Whether each of the various intermediate bases of payment should receive some weight in a reimbursement scheme is an open question. However this is determined, it seems unlikely that it is optimal to put all the weight on one base.

---

[41]For this method to reach a global optimum, the cost function must not have multiple local minima.

References

Ballard, Charles L., and Don Fullerton, "Discretionary Taxes and the Provision of Public Goods," *Journal of Economic Perspectives*, 6, 117–131, 1992.

Ballard, Charles L., and John Goddeeris, "Financing Universal Health Care in the United States: A General Equilibrium Analysis of Efficiency and Distributional Effects," East Lansing: Michigan State University Econometrics and Economic Theory Paper No. 9104, 1993.

Ballard, Charles L., John B. Shoven, and John Whalley, "General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States," *American Economic Review*, 75:128–138, 1985.

Baron, David P., "Design of Regulatory Mechanisms and Institutions," in *Handbook of Industrial Organization*, Vol. 2, Richard Schmalensee and Robert Willig (eds.); Amsterdam: North-Holland, 1989.

Baumol, William J., "Private Affluence, Public Squalor," New York: C.V. Starr Center for Applied Economics (Research Report 92-15), 1992.

Breyer, Friedrich, "The Specification of a Hospital Cost Function," *Journal of Health Economics*, 6:147–158, 1987.

Browning, Edgar, "On the Marginal Welfare Cost of Taxation," *American Economic Review*, 77:11–23, 1987.

Cowing, Thomas G., Alphonse Holtmann, and S. Powers, "Hospital Cost Analysis: A Survey and Evaluation of Recent Studies," in Richard M. Scheffler and Louis F. Rossiter (eds.), *Advances in Health Economics and Health Services Research*, Vol. 4; Westport, CT: JAI Press, 1983.

Custer, William S., James W. Moser, Robert A. Musacchio et al., "The Production of Health Care Services and Changing Hospital Reimbursement: The Role of Hospital-Medical Staff Relationships," *Journal of Health Economics*, 9:167–192, 1990.

Decker, Sandra L., "The Effect of Medicaid on Access to Health Care and Welfare Participation," Ph.D. dissertation, Department of Economics, Harvard University, August 1993, mimeo.

Dranove, David, and Mark A. Satterthwaite, "Monopolistic Competition When Price and Quality Are Imperfectly Observable," *RAND Journal of Economics*, 23:518–534, 1992.

Efron, Bradley, and Carl N. Morris, "Stein's Estimation Rule and Its Competitors: An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68:117–130, 1973.

———, "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, 70:311–319, 1975.

Ellis, Randall P., and Thomas G. McGuire, "Provider Behavior Under Prospective Reimbursement: Cost Sharing and Supply," *Journal of Health Economics*, 5:129–152, 1986.

———, "Optimal Payment Systems for Health Services," *Journal of Health Economics*, 9:375–396, 1990.

Feldstein, Martin S., *Economic Analysis for Health Services Efficiency;* Amsterdam: North-Holland, 1968.

Frech, H.E., and Paul Ginsburg, "Imposed Health Insurance in Monopolistic Markets," *Economic Inquiry*, 13:55–70, 1975.

Fullerton, Don, "Reconciling Recent Estimates of the Marginal Welfare Cost of Taxation," *American Economic Review*, 81:302–308, 1991.

Goodall, Colin, "A Simple Objective Method for Determining a Percent Standard in Mixed Reimbursement Systems," *Journal of Health Economics*, 9:253–271, 1990.

Granneman, Thomas, Randall S. Brown, and Mark V. Pauly, "Estimating Hospital Costs: A Multiple-Output Analysis," *Journal of Health Economics*, 5:107–128, 1986.

Grosskopf, S., and V. Valdmanis, "Measuring Hospital Performance: A Non-Parametric Approach," *Journal of Health Economics*, 6:89–107, 1987.

Haber, Susan G., Jack Zwanziger, Geoffrey M. Anderson et al., "Hospital Expenditures in the United States and Canada: Do Hospital Worker Wages Explain the Difference?" *Journal of Health Economics*, 11:453–465, 1992.

Hadley, Jack, and Stephen Zuckerman, "Determinants of Hospital Costs: Outputs, Inputs, and Regulation in the 1980's," Washington: Urban Institute, Publ. No. 91-10, 1991.

Hendricks, Ann M., "Hospital Wage Gradients Within U.S. Urban Areas," *Journal of Health Economics*, 8:233–246, 1989.

Jensen, Gail A., and Michael A. Morrisey, "Medical Staff Specialty Mix and Hospital Production," *Journal of Health Economics*, 5:253–276, 1986a.

Jensen, Gail A., and Michael A. Morrisey, "The Role of Physicians in Hospital Production," *Review of Economics and Statistics*, 68:432–442, 1986b.

Jorgenson, Dale, "Econometric Methods for Modeling Producer Behavior," in *Handbook of Econometrics*, Vol. 3, Zvi Griliches and Michael D. Intriligator (eds.); Amsterdam: North-Holland, 1986.

Keeler, Emmett B., "What Proportion of Hospital Cost Differences Is Justifiable?" *Journal of Health Economics*, 9:359–365, 1990.

Keeler, Emmett B., Katherine L. Kahn, David Draper et al., "Changes in Sickness at Admission Following the Introduction of the Prospective Payment System," *Journal of the American Medical Association*, 264:1962–1968, 1990.

Laffont, Jean-Jacques, "Toward a Normative Theory of Incentive Contracts Between Government and Private Firms," *Economic Journal*, 97 (Conference), 17–31, 1987.

Laffont, Jean-Jacques, and Jean Tirole, "Using Cost Observation to Regulate Firms," *Journal of Political Economy*, 94:614–641, 1986.

Lau, Lawrence J., "Functional Forms in Econometric Model Building," in *Handbook of Econometrics*, Vol. 3, Zvi Griliches and Michael D. Intriligator (eds.); Amsterdam: North-Holland, 1986.

Lazear, Edward P., and Sherwin Rosen, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 89:841–864, 1981.

Manning, Willard G., et al., "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *American Economic Review*, 77:251–277, 1987.

Marquis, M. Susan, "Cost Sharing and Provider Choice," *Journal of Health Economics*, June 1985.

Nalebuff, Barry J., and Joseph E. Stiglitz, "Information, Competition, and Markets," *American Economic Review*, 73(Supplement):278–283, 1983.

Newhouse, Joseph P., "Pricing and Imperfections in the Medical Care Marketplace," in *Health Economics Worldwide*, Peter Zweifel and H.E. Frech (eds.); Amsterdam: Kluwer, 1991.

———, "Medical Care Costs: How Much Welfare Loss?" *Journal of Economic Perspectives*, 6(3):3–21, 1992.

———, "An Iconoclastic View of Cost Containment," *Health Affairs*, 12(Supplement):152–171, 1993.

Pauly, Mark V., *Doctors and Their Workshops*, Chicago: University of Chicago Press, 1980.

———, "The Economics of Moral Hazard: Comment," *American Economic Review*, 58:531–537, 1968.

Pauly, Mark V., and Michael Redisch, "The Not-for-Profit Hospital as a Physicians' Co-operative," *American Economic Review*, 63:87–99, 1973.

Pope, Gregory C., "Using Hospital-Specific Costs to Improve the Fairness of Prospective Reimbursement," *Journal of Health Economics*, 9:237–251, 1990.

Rogowski, Jeannette R., and Daniel J. Byrne, "Comparison of Alternative Weight Recalibration Methods for Diagnosis-Related Groups," *Health Care Financing Review*, 12(2):87–101, 1990.

Schieber, George J., Jean-Pierre Poullier, and Leslie M. Greenwald, "U.S. Health Expenditure Performance: An International Comparison and Data Update," *Health Care Financing Review*, 13(4):1–87, 1992.

Selden, Thomas M., "A Model of Capitation," *Journal of Health Economics*, 9:397–409, 1990.

Shleifer, Andrei, "A Theory of Yardstick Competition," *RAND Journal of Economics*, 16:319–327, 1985.

Siegel, Carole, Kristine Jones, Eugene Laska et al., "A Risk-Based Prospective Payment System That Integrates Patient, Hospital, and National Costs," *Journal of Health Economics*, 11:1–41, 1992.

Stiglitz, Joseph E., "Incentives and Risk Sharing in Sharecropping," *Review of Economic Studies*, 41:219–255, 1974.

Stuart, Charles, "Welfare Costs per Dollar of Additional Tax Revenue in the United States," *American Economic Review*, 74:352–362, 1984.

Vita, Michael G., "Exploring Hospital Production Relationships with Flexible Functional Forms," *Journal of Health Economics*, 9:1–22, 1990.

Vitaliano, Donald F., "On the Estimation of Hospital Cost Functions," *Journal of Health Economics*, 6:305–318, 1987.

Wagstaff, Adam, "Econometric Studies in Health Economics: A Review of the British Literature," *Journal of Health Economics*, 8:1–51, 1989.

Zeckhauser, Richard J., "Medical Insurance: A Case Study of the Tradeoff Between Risk Spreading and Appropriate Incentives," *Journal of Economic Theory*, 1:10–26, 1970.

Zuckerman, Stephen, Jack Hadley, and Lisa Iezzoni, "Measuring Hospital Efficiency with Frontier Cost Functions," Washington, D.C.: The Urban Institute (Discussion Paper 6172-01), March 1993.