# Efficiency in the Provision of Health Care: An Analysis of Health Maintenance Organizations

James L. Bothwell; Thomas F. Cooley

*Southern Economic Journal*, Vol. 48, No. 4. (Apr., 1982), pp. 970-984.

Stable URL:
http://links.jstor.org/sici?sici=0038-4038%28198204%2948%3A4%3C970%3AEITPOH%3E2.0.CO%3B2-P

*Southern Economic Journal* is currently published by Southern Economic Association.

# Efficiency in the Provision of Health Care: An Analysis of Health Maintenance Organizations*

JAMES L. BOTHWELL
*U. S. General Accounting Office, Washington, D. C. and*
*Duke University, Durham, North Carolina*
THOMAS F. COOLEY
*University of California*
*Santa Barbara, California*

## I. Introduction

Since enactment of the Medicare and Medicaid programs in 1965, a growing public concern over the persistently rising costs of medical care has prompted the imposition of direct regulatory controls over much of the health care industry. The National Health Planning and Resources Development Act of 1974, for example, requires states to enact certificate of need legislation in order to qualify for certain federal subsidies. Under such statutes, all major investment decisions by hospitals, nursing homes and other health care institutions must be approved by designated regulatory agencies. As of 1974, such laws were already in effect in 24 states, and less direct controls over hospital investment were being applied in many others through a provision in the Social Security Amendments of 1972. These amendments also provide for the establishment of quasi-regulatory bodies (Professional Standards Review Organizations) to decide whether the physician and hospital services received by federal beneficiaries under the Medicare, Medicaid, and Maternal and Child Health programs are medically necessary, are being provided in an efficient manner, and are of adequate quality. In addition to these regulatory controls over investment and utilization decisions, several states have imposed traditional public utility rate setting regulation over both hospitals and nursing homes, and some form of universal hospital cost control is a key feature of several prominent national health insurance proposals.[1]

* The views expressed in this paper are strictly those of the authors and do not necessarily represent any official position of the U. S. General Accounting Office. We are grateful to Dennis J. Dugan, H. E. Frech III, Harold Luft, Jack Marshall, Timothy McGuire, Dave O'Neill, Mark Pauly, Michael Redisch, and David Salkever for their comments. An anonymous referee has also contributed many helpful suggestions and comments. Responsibility for errors remains our own. James Bothwell would like to acknowledge the support of the Center for the Study of Business Regulation, Fuqua School of Business, Duke University.

[1] For a collection of papers on the background, extent, and effect of health regulation, see Zubkoff, Raskin, and Hanft [32].

The alternative to direct government regulation of the health care sector is to promote institutional changes which will lead to more efficient resource allocation. One such alternative which has been strongly advocated by many health policy analysts and actively promoted by the federal government is the development of health maintenance organizations (HMOs). Because HMOs provide comprehensive health care on a prepaid basis, it is believed that they will not suffer from the price distortions caused by the third party insurance system and thus will be able to achieve greater efficiency in the production of health care. Whether the development of HMOs will be successful in reducing health care costs depends on the extent to which they respond to incentives by substituting inputs in a cost effective way and the extent to which government policy takes account of the possibility of scale economies in the provision of comprehensive health care. This paper explores one way to address these issues by estimating a joint cost function for HMOs from which substitution elasticities and measures of scale economies are derived. Although the present analysis is limited to a sample of federally qualified HMOs the multi-input multi-output approach we use is generally applicable to other types of health care providers as well.

The next section contains a discussion of the differences in the organizational and incentive structures of HMOs and a brief review of the relevant literature. A model of the HMO as a multiproduct firm is presented in Section III, and the transcendental logarithmic (translog) joint cost function is discussed in Section IV. The data are discussed and the estimates of the cost function are presented in Section V. Sections VI and VII discuss the measures of scale economies and the elasticities of substitution. The results show evidence of the kind of input substitution that would be expected on the basis of the incentive structure implicit in these organizations. There is also evidence of significant increasing returns to scale for HMOs. A short summary and some concluding remarks are contained in the last section.

## II. Organizational Forms, Efficiency, and Economies of Scale

The price mechanism has lost much of its allocative function throughout most of the traditional, fee-for-service sector of the health care industry. Extremely high levels of third party, cost based insurance payments enable medical decision makers to behave as if resources are free. This is especially true for hospital services, where approximately 90 percent of all costs are currently paid by third party insurers. Because HMOs combine the insurance and financing function with the direct provision of health care, it is often argued that HMO decision makers have both the incentive and the ability to react to undistorted prices and allocate resources more efficiently. Unfortunately, the HMO has never been a narrowly defined concept, and differences in organizational structure may adversely affect this incentive.

The cardinal characteristic of an HMO is that it must be an organization that provides comprehensive health care services to voluntarily enrolled, identifiable groups on a prepaid basis with the providers being at some economic risk. The predominant organizational form is the prepaid group practice (PGP), where physicians are members of a closed panel, multispecialty group practice, are paid on either a salary or per member basis, and share centrally located medical facilities. A minor variation of this is the staff model, where physicians are salaried employees of the HMO rather than members of an organizationally separate group practice. The major differences in organizational structure occur in the individual practice associations (IPAs), where the participating physicians are paid on a fee-for-service basis and maintain their own individual office based practices.

The method of remunerating physicians, who have a major role in the decision making process, can affect HMO efficiency. Specifically, the incentive structure for IPA physicians is mixed since their incomes are directly related to the quantities of services they provide. The salaried physicians of PGPs and staff model HMOs, however, have no such countervailing incentive to supply marginal units of care or overutilize other HMO resources.

The available evidence seems to support these views. Several studies have found evidence of both lower hospitalization rates and substantially lower costs for HMO enrollees compared to allegedly similar groups with conventional health insurance who received care from the traditional fee-for-service sector. Furthermore, these differences were greatest for HMOs with salaried physicians.[2] However, whether lower HMO costs may be attributed in part to increases in allocative efficiency stemming from the ability of HMO decision makers to react to relative input costs undistorted by third party payments, requires evidence of substitution between hospital and ambulatory services in the production process.[3] This issue can be addressed by deriving estimates of the elasticities of input substitution directly from an estimated cost function for HMOs. Estimates of the own price elasticities of demand for inputs can be derived as well. Another factor which may affect HMO costs is the realization of economies of scale if they exist in the production of comprehensive prepaid health care. Within the fee-for-service sector, the existence of scale economies has been a much debated issue. On a priori grounds, many economists have expected substantial returns to scale in hospitalization and, because of indivisibilities of capital equipment and the use of allied health personnel, in ambulatory medical care output as well. In the many studies which attempt to measure the extent of such scale economies, the available evidence is quite contradictory and inconclusive on both subjects. Any detailed review of this literature would be well beyond the scope of this paper.[4] However, it is important to note that most cost studies of conventional fee-for-service hospitals and physician office practices employ simple and restrictive functional forms, and, in many cases, omit some inputs and factor prices entirely from the analysis. For example, numerous cost and production function studies of United States hospitals omit any measures of physician input. Thus, if physician input is systematically associated with the size of hospitals, evidence of returns to scale may be illusory.

With regard to studies of ambulatory physician output, it is noteworthy that prior empirical results are based almost entirely on data from fee-for-service, single specialty, private office practices. The production of ambulatory health care by group practice and staff model HMOs, however, is carried on in multispecialty settings where greater degrees of input substitution are possible. Moreover, Roemer and Shonick [28, 297] note that the fee-for-service, private practice physician "typically prescribes treatment for the same

[2] For reviews of this literature see Roemer and Shonick [28], Lewis, Fein and Mechanic [23], Frech and Ginsberg [15] and Luft [22].

[3] This is not to say that similar substitution is not possible in the traditional fee-for-service sector, but only that the relative input costs facing HMO decision makers will not be subject to the distortions caused by differential third party payments. Davis and Russell [8], for example, provide some evidence of substitution between outpatient and inpatient care in conventional hospitals.

[4] For reviews of this literature see Newhouse [26], Mann and Yett [24] and Berki [3].

patient in the office and in the hospital . . . (and) can ostensibly increase his efficiency of practice by hospitalizing patients, passing along the heavy diagnostic work to the hospital and the expense to the insurance plans." Since larger group practices generally have a wider range of diagnostic and therapeutic services available in their own offices, empirical studies based on fee-for-service practice which employ only an aggregate measure of physician ambulatory output and ignore the possibility of joint hospital production may give misleading results.

Roemer and Shonick also suggest that some economies may be peculiar to the PGP mode of organization, although there is little empirical support for this contention because of the paucity of data on PGPs. The market dominance of a few very large size HMOs, however, is undisputable. As of 1978 there were an estimated 199 operational HMOs with a combined enrollment of 7.4 million persons.[5] In spite of the fairly large number of prepaid plans, 12 HMOs with 100,000 or more members accounted for approximately 70 percent of total membership; and 5 of these were Kaiser Foundation Health plans with a combined enrollment of 3.4 million members.

The issue of scale economies in the production of comprehensive prepaid medical care is a crucial one since almost all of the comparative studies showing substantially lower costs for HMO members have been based on evidence from one or two HMOs in this small, unrepresentative set of very large organizations. If increasing returns to scale are significant, this implies that these previous findings may be in part due to size and thus may not be attributable to HMOs in general. This has serious implications for the success of the current federal HMO strategy which, under the provisions of the Health Maintenance Organization Act of 1973, is primarily providing financial assistance for the establishment and operation of much smaller HMOs, some of which are located in areas with a potentially limited demand for their services.[6]

Fortunately, enough data are available for us to test for economies of scale on a sample of HMOs ranging in size from 1,131 to 37,000 members. More importantly, because HMOs provide comprehensive care and must bear all the costs of treatment regardless of where they are incurred, we are able to derive these estimates from a properly specified mutli-input, multi-output model which explicitly recognizes the nature of joint production in medical care and avoids some of the shortcomings of previous studies.

## III. Health Maintenance Organizations as Multiproduct Firms

Health maintenance organizations provide virtually all of the services that are commonly provided by the fee-for-service sector, with the distinguishing characteristic that the services are supplied or contracted for by one organization. Thus, it is appropriate to regard them as multiproduct firms producing a vector of outputs from a vector of inputs.

[5] These statistics are compiled by InterStudy [18].

[6] The HMO Act, as amended, provides for grants and federal loan guarantees for HMO feasibility studies, planning projects, and initial development costs. Federal loans and loan guarantees are also available for up to $2.5 million per HMO for the acquisition or construction of ambulatory health care facilities. An additional $2.5 million in loans and guarantees is available to cover an HMO's operating deficits during the first five years. Total grant assistance for all purposes for fiscal years 1975–78 was $74.5 million. As of February 1978 federally qualified HMOs have received direct federal loans totaling $119 million and guarantees of $3.5 million. The average size of the 48 operational HMOs receiving federal financial assistance as of March 1978 was 14,023 members.

They are characterized by common costs where these are defined to be the costs of common inputs which are utilized by more than one output. A multiproduct production process can be represented by the product transformation function:

$$F(Y_1, \ldots., Y_m; X_1, \ldots., X_n) = 0 \tag{1}$$

where the $Y_i$ represent the outputs and the $X_j$ represent the inputs.

The theory of duality between cost and production implies that there will exist a dual cost function to the product transformation function (1) if the following assumptions hold true:[7]

i. the firm pursues cost minimizing behavior;
ii. the firm has no control over input prices;
iii. the product transformation surface satisfies regularity conditions (i.e., convex isoquants).

The dual cost function will have the form

$$C = C(Y_1, \ldots., Y_m; P_1, \ldots., P_n) = 0 \tag{2}$$

where the $P_j$ represent the prices of the inputs $X_j$. The cost function, then, is a complete description of production technology and contains virtually all of the information that the product transformation function contains.

The cost function (2) has the properties:

i. $C$ is increasing in $Y_i$ and $P_j$;
ii. $C$ is linear homogeneous in $P_j$;
iii. $C$ is concave in $P_j$;
iv. $\partial C/\partial P_j = X_j$ (Shephard's Lemma).

While the description of the production process embodied in equation (1) is quite appropriate for HMOs, the existence of a dual cost function depends on the validity of assumptions i.–iii. There is no reason to assume that the production technology of HMOs would be irregular or to assume that they would have control over input prices. Newhouse [25] has suggested, however, that the assumption of cost minimizing behavior is questionable for nonprofit institutions that receive cost reimbursement payments from third parties. Although this may be true of conventional hospitals, it is not the case for the HMO which must compete with conventional health care providers and insurers as well as other HMOs.[8] Moreover, some HMOs are for profit enterprises, while others like the noted Kaiser plans reinforce their overall cost minimizing incentive by instituting "profit" sharing plans with their physicians. Thus, all three of the necessary assumptions appear quite valid for the purposes of this analysis.

## IV. The Translog Joint Cost Function

Historically, many empirical studies of cost functions have employed functional forms which imply strong restrictions on the type of economic behavior they represent. Duality theory suggests that the form of the cost function has implications for the nature of the underlying production process [16]. More recently, the translog functional form has

[7] See Diewert [9; 10] for further details.

[8] Evidence of the competitive impact of HMOs has been presented in recent reports by the Federal Trade Commission [12] and InterStudy [19].

become increasingly popular as a representation of cost functions because it enables one to model costs without unnecessary prior restrictions on the production process and restrictive prior assumptions about the substitutability of inputs. The translog function is quadratic in logarithms and is one of the family of second-order Taylor series approximations to an arbitrary cost function. For the multiple output firm, the translog function takes the form

$$\begin{aligned}\log C = \alpha_0 + \sum_{i=1}^{m} \alpha_i \log Y_i + \sum_{j=1}^{n} \beta_j \log P_j \\ + ½ \sum_{i=1}^{m} \sum_{l=1}^{m} \delta_{il} \log Y_i \log Y_l + ½ \sum_{j=1}^{n} \sum_{k=1}^{n} \gamma_{jk} \log P_j \log P_k \\ + \sum_{i=1}^{m} \sum_{j=1}^{n} \rho_{ij} \log Y_i \log P_j \qquad (3)\end{aligned}$$

where the $\alpha_i$, $\beta_j$, $\delta_{il}$, $\gamma_{ij}$, and $\rho_{ij}$ are parameters to be estimated. Shephard's Lemma ($\partial C/\partial P_j = X_j$) implies

$$\partial \log C/\partial \log P_j = P_jX_j/C = M_j \qquad (4)$$

where $M_j$ is the cost share of the $j$th input. Applying (4) to (3) yields the system of cost-share equations

$$M_j = \beta_j + \sum_{k} \gamma_{jk} \log P_k + \sum_{i} \rho_{ij} \log Y_i, j = 1, \dots, n. \qquad (5)$$

The system of equations (3), (5) is the cost system to be estimated. For a production process with $m$ outputs and $n$ inputs there is a total of $m + n + m^2 + n^2 + mn$ parameters. The fact that the function is a second-order approximation implies symmetry of the form $\delta_{il} = \delta_{li}$ and $\gamma_{jk} = \gamma_{kj}$. Further, since the $M_j$ are cost shares, $\Sigma_{i=1}{}^n M_j = 1$ which implies $\Sigma_j\beta_j = 1$, $\Sigma_j\gamma_{jk} = 0$ and $\Sigma_j\rho_{ij} = 0$. Finally, the fact that cost functions must exhibit homogeneity of degree +1 in input prices implies $\Sigma_k\gamma_{jk} = 0$. This reduces the number of free parameters to $mn + (m+1)(m/2) + (n+1)(n/2)$.

By imposing parameter restrictions on the translog cost function, it is possible to test whether or not the technology exhibits constant returns to scale and whether or not the vector of outputs is separable from the vector of inputs. Constant returns to scale imply the restrictions

$$\sum_{i=1}^{m} \alpha_i = 1, \sum_{i=1}^{m} \delta_{il} = 0, \sum_{i=1}^{m} \rho_{ij} = 0 \qquad (6)$$

in addition to those already discussed, but only $n-1$ of the last set of the restrictions are independent since $\Sigma_j\rho_{ij} = 0$ has already been imposed.

In testing for separability we consider only strong separability on the translog cost function itself rather than on the underlying cost function which is being approximated. Brown, Caves and Christensen [5] show that

$$\rho_{ij} = 0, i=1, \dots, m, j=1, \dots, n, \qquad (7)$$

is a sufficient condition for strong separability.

## V. Estimates of a Translog Joint Cost Function for HMOs

### *The Data*

The Health Maintenance Organization Act was passed in 1973 to provide federal support for HMO growth and development. HMOs qualified to receive assistance under the provisions of the Act must offer specified, comprehensive services, have community rated premiums, institute quality assurance and utilization review programs, charge only nominal coinsurance rates, and strictly limit the amount of reinsurance, if any. In addition, each qualified and operational HMO is required to provide detailed data on costs, membership, services provided, and a variety of other aspects of its operation. The HMO National Data Reporting Requirements were developed and implemented by the Department of Health, Education, and Welfare to collect these data on uniform quarterly, semiannual, and annual reports.

As of the fourth quarter of 1977, there were 36 qualified and operational HMOs for which data were available through this reporting system. Of these 36 HMOs, ten were individual practice associations (IPAs), eight were prepaid group practices, and eighteen were staff model HMOs. Because of major differences in organization, incentive structures, and reporting requirements, we eliminated the IPAs from the data set. The study thus utilizes quarterly observations for the period from the first quarter of 1976 to the fourth quarter of 1977 on federally qualified prepaid group practice and staff model HMOs. It was also necessary to eliminate all observations for which some of the data were not available. This reduced the number of observations from 208 to 106 and the number of HMOs in the sample from 26 to 20. These remaining HMOs ranged in size from 1,131 to over 37,000 members as of December 1977. The oldest of these was established in 1971, while the newest HMO in the sample became operational in 1977.

The definition of output has always been problematic in empirical studies of the health care industry. Although one can conceptualize final output as the improvement or maintenance of the health status of individuals, the lack of suitable indices of health negates this approach. Therefore, in this study we use measures of three intermediate outputs: ambulatory encounters with physicians (*A1*), ambulatory encounters with allied health care professionals (*A2*), and hospital discharges (*HD*). Four inputs are distinguished: administrative services (*AD*), hospital services (*HS*), medical professional staff services (*ME*), and health center services (*HC*). Essentially, the latter are the capital expenses of maintaining a health center.[9]

[9] To obtain estimates of implicit input prices, we divided the aggregate expenses of these services by ordinal proxies measuring aggregate input usage. Specifically, we defined the following input prices:

administrative services price = (health plan administration expense)/(member months)
hospital services price = (hospitalization expense)/(hospital days)
medical professional staff services price = (medical group expense for direct service and outside referrals + special services expense)/(full-time equivalent medical care personnel including physicians, physician extenders, nurses, optometrists, podiatrists, mental health care providers, dental health care providers, and other direct health care providers)
health center services price = (health center expense + interest expense on loans)/(member months)

Since we are forced by the data limitations to use proxy quantities to obtain these prices, there is a possibility that they will still contain some endogenous component.

## *Estimates*

Two modifications are made in the basic translog framework to account for special characteristics of the data. First, many of the HMOs in the sample became operational during the period we are studying in response to the incentives offered by the HMO Act of 1973. Others became qualified for federal assistance under the Act during this period, but had been operational prior to 1973. Variation in the length of time HMOs have been in operation may have important effects on costs. On the one hand, newly formed firms may experience some inefficiencies during the first few quarters of operation which diminish over time. On the other hand, many new medical care techniques and technologies are cost increasing.[10] To control for these effects, we introduce a variable equal to the number of quarters in operation of the HMO. It may be interpreted as a measure of Hicks neutral technical change. Because the net effect could either be cost decreasing or cost increasing, however, we have no a priori expectation concerning the sign of its coefficient.

The second modification is to control for variation in the length of stay for hospitalized patients. Since longer hospital stays require more routine nursing and "hotel" type services, and usually entail a higher cost per discharge, we expect a positive effect on costs. In addition, as a major study by Anderson and Sheatsley [1] shows, longer than average hospital stays are generally associated with more serious types of illnesses, which require more extensive diagnostic and therapeutic services. Of course, it would have been preferable to also control for variation in casemix directly by including as independent variables the proportion of hospitalized patients in specific diagnostic categories, or an index based on such information. Unfortunately the only such data available were for inpatient discharges based on the following very broad categories: medical/surgical; obstetrical; newborns in hospital; mental health; all others. These breakdowns, however, were much too aggregated to provide any meaningful information about differences in casemix. However, sufficient data on both length of stay and diagnostic casemix for conventional hospitals were available in studies by Evans [11] and Pauly [27]. Evans reported that the average length of stay was highly collinear with diagnostic casemix variables and had a strongly positive and significant effect on cost per case. Pauly found that when the average length of stay was added to a regression of total hospital cost on variables measuring diagnostic casemix, physician staff characteristics and hospital characteristics, it caused the casemix variables to become statistically insignificant. The esimated coefficient on length of stay, however, was positive and significant. Thus, it appears from both of these studies, which were based on entirely different data sets, that an average length of stay variable can capture much of the same information available from diagnostic casemix variables.[11]

The assumed form of the cost function is:

$$C = f(T,LOS)\ C^*\ (Y_1, \ldots, Y_m; P_1, \ldots, P_n) \tag{8}$$

[10] See Feldstein [13] and Russell [29] for evidence of this.

[11] In the absence of any meaningful data on diagnostic casemix, previous studies of hospital cost functions have attempted to standardize output either by assuming that casemix is constant for a hospital over a short period of time, or by grouping hospitals on the basis of the facilities and services available. See, for example, Lave and Lave [21], Carr and Feldstein [6], Cohen [7], and Francisco [14]. Neither of these procedures is relevant in this instance, however, since we use, at most, two years of data per HMO, and all federally qualified HMOs are required to offer the same benefit package. Moreover, the very detailed data on the age-sex composition of HMO membership which were available showed very little variation which indicates a great degree of homogeneity in the enrolled populations.

where $T$ is time in operation, $LOS$ is the average length of inpatient stay for the HMO for each quarter, and $C^*$ is the basic cost function.[12] The function $f$ is assumed to be exponential so that $T$ and $LOS$ enter the translog form additively. Because the data are time series of cross sections with unequal observations on the cross section, we initially specified the error process to contain both HMO specific and time specific components as well as a general component:

$$\varepsilon_{it} = u_i + v_t + w_{it} \qquad i = 1, \ldots, 20, t = 1, \ldots, 8), \tag{9}$$

where

$$u_i \sim N(0, \sigma_u^2),$$
$$v_t \sim N(0, \sigma_v^2),$$
$$w_{it} \sim N(0, \sigma_w^2).$$

The variance components are assumed independent of one another as well as temporally and cross sectionally independent. Estimates of the cross section and time specific variance components were not significantly different from zero. Consequently, in all subsequent estimation, we assumed only a common variance.

The system of equations estimated includes the modified translog cost function and $n - 1$ of the cost-share equations (5). Classical additive disturbances are assumed for all equations. Since the cost shares must sum to 1, the disturbances to (5) must sum identically to zero. As this would imply a singular covariance matrix for the system of equations one of the share equations must be eliminated. The share equation corresponding to health center services (*HC*) was eliminated, and the systems estimation procedure proposed by Zellner [31] was applied iteratively until covergence was achieved.[13]

Table I presents estimates of the translog cost function with no restrictions other than those implied by linear homogeneity and the share equations, with parameter restrictions which imply constant returns to scale, and with parameter restrictions which imply separability of the outputs. The subscripts of the parameters represent the two controlling variables, the three outputs, and the four inputs defined earlier.

It is interesting to note that in the unrestricted form of the translog function the coefficients of the controlling variables are both positive and significant. Technical change, represented by $\mu_T$, increases total cost, while, as expected, the positive sign on length of stay, $\mu_{LOS}$, indicates that HMOs with longer average stays have higher total costs. To test the hypotheses of constant returns to scale and separability we compute the likelihood ratio statistic

$$-2 \ln \lambda = n(\ln|\hat{\Sigma}_\gamma| - \ln|\hat{\Sigma}_u|) \tag{10}$$

where $|\hat{\Sigma}_\gamma|$ is the determinant of the estimated covariance matrix with the restrictions imposed and $|\hat{\Sigma}_u|$ is the determinant of the covariance matrix of the unrestricted system.

[12] This specification, of course, treats length of stay as exogenous. There are two reasons why we feel this is a plausible assumption in the case of group practice and staff model HMOs. First, staff physicians are salaried and thus have no financial incentives to extend the length of hospital stays. Second, all length of stay decisions in HMOs are routinely scrutinized by utilization review panels. Thus, for a given diagnosis, differences in the length of inpatient stays for HMO members are more likely to reflect exogenous differences in case complexity and severity than disparities in the discharge practices of many individual physicians.

[13] Kmenta and Gilbert [20] show that iterations of Zellner's estimator will converge to maximum likelihood estimates (if they converge), while Barten [2] demonstrates that such parameter estimates are independent of which share equation is dropped from the system.

**Table I.** Estimates of Translog Cost Functions

| | Unrestricted | | Constant returns to scale | | Separable | |
|---|---|---|---|---|---|---|
| Variable | Estimated Coefficient | Standard Deviation | Estimated Coefficient | Standard Deviation | Estimated Coefficient | Standard Deviation |
| $\alpha_0$ | 4.078 | 2.380 | 8.237 | 1.209 | 7.999 | 1.056 |
| $\mu_T$ | 0.012 | 0.003 | 0.002 | 0.003 | −0.003 | 0.003 |
| $\mu_{LOS}$ | 0.061 | 0.012 | 0.077 | 0.014 | 0.081 | 0.014 |
| $\alpha_{A1}$ | 0.337 | 0.879 | −1.830 | 0.527 | −1.488 | 0.534 |
| $\alpha_{A2}$ | −0.301 | 0.324 | 0.370 | 0.252 | 0.213 | 0.242 |
| $\alpha_{HD}$ | 1.481 | 0.817 | 2.460 | 0.629 | 2.275 | 0.613 |
| $\beta_{AD}$ | 0.778 | 0.074 | 0.944 | 0.062 | 0.866 | 0.044 |
| $\beta_{HS}$ | 0.133 | 0.084 | 0.151 | 0.070 | 0.100 | 0.044 |
| $\beta_{ME}$ | −0.452 | 0.109 | −0.609 | 0.102 | −0.531 | 0.064 |
| $\beta_{HC}$ | 0.541 | 0.085 | 0.515 | 0.076 | 0.565 | 0.053 |
| $\delta_{A1,A1}$ | 0.856 | 0.183 | 0.620 | 0.132 | 0.588 | 0.148 |
| $\delta_{A1,HD}$ | −0.199 | 0.158 | −0.507 | 0.137 | −0.520 | 0.147 |
| $\delta_{HD,HD}$ | 0.252 | 0.158 | 0.454 | 0.172 | 0.511 | 0.176 |
| $\delta_{A1,A2}$ | 0.005 | 0.063 | −0.113 | 0.053 | −0.068 | 0.053 |
| $\delta_{A2,A2}$ | 0.060 | 0.026 | 0.060 | 0.030 | 0.058 | 0.030 |
| $\delta_{A2,HD}$ | −0.16 | 0.065 | 0.053 | 0.070 | 0.009 | 0.071 |
| $\gamma_{AD,AD}$ | 0.174 | 0.006 | 0.159 | 0.004 | 0.171 | 0.005 |
| $\gamma_{AD,HS}$ | −0.060 | 0.004 | −0.065 | 0.004 | −0.064 | 0.004 |
| $\gamma_{AD,ME}$ | −0.062 | 0.006 | −0.049 | 0.005 | −0.058 | 0.006 |
| $\gamma_{AD,HC}$ | −0.052 | 0.006 | −0.045 | 0.004 | −0.049 | 0.005 |
| $\gamma_{HS,HS}$ | 0.113 | 0.009 | 0.124 | 0.009 | 0.104 | 0.009 |
| $\gamma_{HS,ME}$ | −0.022 | 0.008 | −0.029 | 0.007 | −0.008 | 0.007 |
| $\gamma_{HS,HC}$ | −0.031 | 0.006 | −0.031 | 0.006 | −0.032 | 0.006 |
| $\gamma_{ME,ME}$ | 0.121 | 0.011 | 0.106 | 0.009 | 0.098 | 0.010 |
| $\gamma_{ME,HC}$ | −0.037 | 0.006 | −0.028 | 0.006 | −0.032 | 0.006 |
| $\gamma_{HC,HC}$ | 0.119 | 0.010 | 0.104 | 0.010 | 0.113 | 0.009 |
| $\rho_{A1,AD}$ | −0.004 | 0.012 | −0.023 | 0.010 | $\bar{R}^2 = .98$ | |
| $\rho_{A1,HS}$ | −0.026 | 0.015 | −0.027 | 0.013 | | |
| $\rho_{A1,ME}$ | 0.035 | 0.019 | 0.060 | 0.017 | | |
| $\rho_{A1,HC}$ | −0.005 | 0.012 | −0.010 | 0.011 | | |
| $\rho_{HD,AD}$ | 0.040 | 0.012 | 0.025 | 0.012 | | |
| $\rho_{HD,HS}$ | 0.043 | 0.016 | 0.044 | 0.015 | | |
| $\rho_{HD,ME}$ | −0.093 | 0.021 | −0.074 | 0.020 | | |
| $\rho_{HD,HC}$ | 0.010 | 0.013 | 0.004 | 0.013 | | |
| $\rho_{A2,AD}$ | −0.004 | 0.005 | −0.002 | 0.005 | | |
| $\rho_{A2,HS}$ | −0.016 | 0.007 | −0.017 | 0.006 | | |
| $\rho_{A2,ME}$ | 0.016 | 0.009 | 0.014 | 0.008 | | |
| $\rho_{A2,HC}$ | 0.004 | 0.006 | 0.006 | 0.005 | | |
| | $\bar{R}^2 = .99$ | | $\bar{R}^2 = .98$ | | | |

This statistic is distributed asymptotically as chi-square with degrees of freedom equal to number of restrictions being tested. The values of the test statistic are presented in Table II. Both hypotheses are soundly rejected, and so the unrestricted translog form must be used. Having rejected the possibility of constant returns to scale, we turn to the problem of measuring returns to scale for HMOs.

**Table II.** Tests for Constant Returns to Scale and Separability

| | $-2 \ln \lambda$ | Number of restrictions $\gamma$ | Critical $\chi^2_\gamma$ at 0.05 level |
|---|---|---|---|
| Constant Returns to Scale | 38.859 | 7 | 14.067 |
| Separable | 57.146 | 9 | 16.919 |

## VI. Economies of Scale

The measurement of returns to scale is more complex for firms which produce multiple outputs because it is necessary to distinguish between returns to scale in some overall sense where all outputs are expanded, and returns to scale with respect to a particular output. Hanoch [17] suggests that it is most appropriate to measure overall returns to scale along an expansion path where all outputs are increased proportionately.

If we assume that all outputs are increased in proportion

$$dY_i/Y_i = d \log Y_i = \lambda, \tag{11}$$

then the measure of scale economies (*SE*) is

$$SE = d \log C/\lambda = \sum_{i=1}^{m} (\partial \log C/\partial \log Y_i) \ . \tag{12}$$

If $SE > 1$ there are overall decreasing returns to scale; if $SE = 1$ the technology exhibits constant returns to scale (which has been ruled out by the test presented in the previous section); and if $SE < 1$ there are overall increasing returns to scale.

For a single output we can consider the elasticity of cost with respect to a single output, all other outputs held constant:

$$SE(i) = (\partial \log C/\partial \log Y_i)|(Y_j), j \neq i \text{ constant} = \partial \log C/\partial \log Y_i \ . \tag{13}$$

If $SE(i) > 1$ there are decreasing returns to scale with respect to the *i*th output; if $SE(i) < 1$ there are increasing returns to scale; and if $SE(i) = 1$ there are constant returns to scale. Clearly, it is possible to have the overall measure of returns to scale, *SE*, indicate decreasing returns to scale while each of the individual returns to scale indicates increasing returns to scale.

An alternative indicator of returns to scale with respect to a single output is the change in incremental costs

$$(\partial^2 C/\partial Y_i^2)|Y_j, \qquad j \neq i \text{ constant.}$$

Decreasing marginal cost ($\partial^2 C/\partial Y_i^2 < 0$) should be indicative of increasing returns to scale with respect to the *i*th output; but, of course, it is possible to observe decreasing marginal cost with respect to each output, and at the same time have overall decreasing returns to scale ($SE > 1$). For the translog function marginal costs are defined as

$$\partial C/\partial Y_j = (\partial \log C/\partial \log Y_j)(\hat{C}/Y_j) = \left(\alpha_j + \sum_{j=1}^{m} \delta_{ij} \log Y_j + \sum_{i=1}^{m} \rho_{ij} \log P_j\right)(\hat{C}/Y_j) \tag{14}$$

where $\hat{C}$ is the fitted value of total costs. It can also be shown that

$$\partial^2 C/\partial Y_i^2 = (\hat{C}/Y_i^2)(\partial^2 \log C/\partial \log Y_i^2 + (\partial \log C/\partial \log Y_i)(\partial \log C/\partial \log Y_i - 1)). \quad (15)$$

Table III presents estimates of overall economies of scale (*SE*), the output cost elasticities (*SEA1*, *SEA2*, *SEHD*) for each output, the marginal costs (*MCA1*, *MCA2*, *MCHD*) for each output, and the derivatives of marginal costs (*DMCA*1, *DMCA*2, *DMCHD*) for each output by HMO.

The measure of overall returns to scale and the measures of returns to scale for individual outputs give quite consistent and unambiguous evidence of increasing returns to scale. The overall measures are less than one for 19 of the 20 HMOs, indicating that total costs are increasing at a decreasing rate along the expansion paths of these firms. For all 20 HMOs, the individual cost elasticities are substantially less than one, giving a clear indication that these HMOs are operating well within the region of increasing returns to scale with respect to the three outputs individually. The marginal costs of ambulatory encounters with physicians are decreasing for all HMOs in the sample, and the marginal costs of ambulatory encounters with allied health professionals are decreasing for 18 of the 20 HMOs. The marginal costs of hospital discharges are increasing for all HMOs, but, of course, this is not inconsistent with increasing returns.

For all 20 HMOs, the marginal cost of a hospital stay is substantially higher than the marginal cost of either type of ambulatory encounter. Somewhat surprising is the fact that for only 4 of the 20 HMOs is the estimated marginal cost of an ambulatory encounter with a physician higher than the marginal cost of an ambulatory encounter with a non-physician health professional. The other HMOs show lower, although in most cases not substantial-

**Table III.** Scale Economies and Marginal Costs by HMO

| HMO | *SE* | *SEA1* | *SEA2* | *SEHD* | *MCA1* | *MCA2* | *MCHD* | *DMCA1* | *DMCA2* | *DMCHD* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .847 | .172 | .256 | .420 | 28.41 | 52.59 | 2240.00 | −.001 | −.004 | .612 |
| 2 | .788 | .191 | .276 | .321 | 26.26 | 62.06 | 1672.29 | −.001 | −.003 | .480 |
| 3 | .892 | .403 | .064 | .425 | 117.12 | 81.07 | 5047.42 | −.049 | 2.892 | 7.625 |
| 4 | .814 | .186 | .291 | .337 | 22.00 | 43.12 | 1690.84 | −.000 | −.001 | .434 |
| 5 | .845 | .260 | .191 | .394 | 48.80 | 69.53 | 1972.47 | −.005 | −.016 | .558 |
| 6 | .832 | .372 | .181 | .280 | 38.79 | 48.01 | 1520.06 | −.004 | −.014 | 3.930 |
| 7 | .842 | .479 | .228 | .135 | 47.08 | 48.19 | 1924.02 | −.007 | −.008 | .115 |
| 8 | .850 | .182 | .254 | .414 | 21.08 | 40.78 | 1853.03 | −.001 | −.002 | .268 |
| 9 | .871 | .154 | .195 | .522 | 24.27 | 53.56 | 1675.36 | −.001 | −.008 | .094 |
| 10 | .720 | .800 | .218 | −.298 | 51.63 | 59.70 | 601.62 | −.005 | −.039 | 144.392 |
| 11 | .678 | .331 | .100 | .248 | 35.35 | 187.01 | 1146.66 | −.002 | −.132 | 2.120 |
| 12 | .836 | .231 | .239 | .367 | 33.65 | 44.51 | 1468.72 | −.002 | −.005 | .448 |
| 13 | .836 | .436 | .188 | .212 | 115.77 | 105.55 | 3028.55 | −.014 | −.035 | 41.385 |
| 14 | .856 | .290 | .181 | .385 | 37.39 | 48.34 | 1705.72 | −.004 | −.012 | 1.252 |
| 15 | .905 | .240 | .183 | .482 | 37.45 | 48.83 | 2180.46 | −.003 | −.008 | .058 |
| 16 | .749 | .260 | .092 | .397 | 23.15 | 104.62 | 1490.11 | −.001 | −.022 | .760 |
| 17 | .823 | .309 | .214 | .301 | 46.71 | 51.35 | 1659.57 | −.004 | −.008 | 1.779 |
| 18 | 1.002 | .344 | .171 | .487 | 134.57 | 38.90 | 3673.39 | −.076 | −.012 | .748 |
| 19 | .814 | .399 | .093 | .321 | 64.60 | 78.50 | 2600.88 | −.010 | .196 | 4.562 |
| 20 | .896 | .626 | .126 | .144 | 178.45 | 87.59 | 3621.80 | −.039 | −.079 | 261.112 |
| Mean | .835 | .333 | .187 | .345 | 56.63 | 67.69 | 2138.65 | −.012 | .134 | 23.637 |

ly lower, marginal costs of physician encounters. This could be the result of HMOs using allied health professional for more capital intensive procedures previously performed by physicians, such as the administration of diagnostic tests or routine therapeutic treatments. The mean values for the marginal costs of the three outputs are \$56.63, \$67.69, and \$2138.65, respectively.[14]

## VII. Elasticities of Substitution

Uzawa [30] has demonstrated that elasticities of substitution can be computed from the cost function as

$$\sigma_{ij} = C((\partial^2 C/\partial P_i \partial P_j)/((\partial C/\partial P_i)(\partial C/\partial P_j))) \ . \tag{16}$$

In the translog cost function these become

$$\begin{aligned} \hat{\sigma}_{ij} &= (\hat{\gamma}_{ij} + \hat{M}_i\hat{M}_j)/\hat{M}_i\hat{M}_j, \ i \neq j, \\ \hat{\sigma}_{ii} &= (\hat{\gamma}_{ii} + \hat{M}_i(\hat{M}_i - 1))/\hat{M}_i^2, \end{aligned} \tag{17}$$

where the $\hat{M}$ are the fitted values of the cost share equations. Berndt and Wood [4] have shown that the own price elasticities of demand can be computed as:

$$\hat{\eta}_{ii} = \hat{M}_i\hat{\sigma}_{ii} = (\hat{\gamma}_{ii} + \hat{M}_i(\hat{M}_i - 1))/\hat{M}_i \tag{18}$$

Concavity of the cost function in input prices requires that $\sigma_{ii} < 0$ for each factor input. A sufficient condition for concavity is that the bordered Hessian be negative semi-definite. In our sample, two firms failed to satisfy either the necessary or the sufficient condition.

Table IV presents estimates of the own price elasticities and the elasticities of substitution. These are the averages for all firms in all time periods which satisfy the concavity condition. As expected, the own price elasticities reveal that demand for all inputs are inelastic with the demand for administrative services being the least elastic. The elasticities of substitution reveal that administrative services are complements to all the other inputs, but that there is substitution between hospital services and medical staff services, between hospital services and health center services, and between medical staff services and health center services.

## VIII. Summary and Concluding Remarks

This paper reports on the application of a translog joint cost function to federally qualified health maintenance organizations. This approach is desirable because it explicitly recognizes the multiproduct nature of health care services and allows one to model costs without unnecessary prior restrictions on the production process and restrictive prior assumptions about the substitutability of inputs. The estimated joint cost function not only provides information about economies of scale, but also allows one to derive estimates of

[14] A few of the individual marginal costs seem large, but these generally are for HMOs which have just started operation and have had relatively few patients. These marginal costs are also all inclusive and, consequently, contain items which are not likely to be included in other estimates of medical care costs.

**Table IV.** Estimated Elasticities of Demand and Substitution

| | Administrative Services | | | Hospital Services | | Medical Staff Services |
|---|---|---|---|---|---|---|
| | Hospital Services | Medical Staff Services | Health Center Services | Medical Staff Services | Health Center Services | Health Center Services |
| Elasticity of Substitution | −0.636 | −0.150 | −0.138 | 0.614 | 0.805 | 0.638 |

| | Administrative Services | Hospital Services | Medical Staff Services | Health Center Services |
|---|---|---|---|---|
| Own Price Elasticity of Demand | −0.104 | −0.287 | −0.283 | −.253 |

the marginal cost of each joint product, the elasticities of substitution between inputs, and the own price elasticities of demand for each input.

The principal results of our analysis are quite plausible. The demands for all inputs are inelastic, and the mean values for the marginal costs of all outputs are within reasonable bounds. Second, there is clear evidence of the kind of input substitution among hospital services, medical services and ambulatory health care center services, that one would expect on the basis of the incentive structure implicit in HMOs. Third, the results show evidence of significant increasing returns to scale for all the HMOs in our sample, which range in size from 1,130 to 37,000 members.

Although these findings have implications for the current federal strategy toward HMO development, they should be treated with caution because the data have many limitations and some of the definitions employed yield only proxies for the appropriate conceptual variables. Because of these considerations, it is perhaps premature to recommend any broad changes in policy or to make any generalizations concerning the structure of production for all HMOs based on these results. Nevertheless, given the ubiquitous nature of joint production in health care, the multi-input multi-output approach we use can have much wider applications to other types of providers. Certainly the results of such analyses, if based on properly specified joint cost or production functions, could make significant contributions to the formulation of public policy toward the health care sector in general.

## References

1. Anderson, Odin W. and P. Sheatsley. *Hospital Use—A Survey of Patient and Physician Decisions.* Chicago: University of Chicago Press, 1967.

2. Barten, A. P., "Maximum Likelihood Estimation of a Complete System of Demand Equations." *European Economic Review,* Fall 1969, 7–78.

3. Berki, Sylvester E. *Hospital Economics.* Lexington: Lexington Books, 1972.

4. Berndt, E. R. and D. O. Wood, "Technology, Prices, and the Derived Demand for Energy." *Review of Economics and Statistics,* August 1975, 259–68.

5. Brown, Randall S., Douglas W. Caves, and Laurits R. Christensen, "Modelling the Structure of Cost and Production for Multiproduct Firms." *Southern Economic Journal,* July 1979, 256–73.

6. Carr, W. John and Paul J. Feldstein, "The Relationship of Cost to Hospital Size." *Inquiry,* June 1967, 45–65

7. Cohen, Harold A., "Variations In Cost Among Hospitals of Different Sizes." *Southern Economic Journal,* January 1967, 355–66.

8. Davis, Karen and Louise B. Russell, "The Substitution of Hospital Outpatient Care for Inpatient Care." *Review of Economics and Statistics,* May 1972, 109–20.

9. Diewert, W. Erwin, "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function." *Journal of Political Economy,* May/June 1971, 481–507.

10. ———. "Applications of Duality Theory," in *Frontiers of Quantitative Economics,* edited by Michael D. Intriligator and David A. Kendrick. Amsterdam: North-Holland Publishing Company, 1974.

11. Evans, Robert G. " 'Behavioral' Cost Functions for Hospitals." *Canadian Journal of Economics,* May 1971, 198–215.

12. Federal Trade Commission, Bureau of Economics. *Staff Report on the Health Maintenance Organization and Its Effects on Competition.* Washington: Federal Trade Commission, July 1977.

13. Feldstein, Martin S. *The Rising Cost of Hospital Care.* Washington: Information Resource Press, 1971.

14. Francisco, Edgar. "Analysis of Cost Variation Among Short Term General Hospitals," in *Empirical Studies in Health Economics,* edited by Herbert Klarman. Baltimore: Johns Hopkins Press, 1970.

15. Frech, H. E. and Paul B. Ginsburg. *Public Insurance in Private Medical Markets.* Washington, D.C.: American Enterprise Institute, 1978.

16. Hall, Robert E., "The Specification of Technologies with Several Kinds of Outputs." *Journal of Political Economy,* July/August 1973, 878–92.

17. Hanoch, Giora, "The Elasticity of Scale and the Shape of Average Costs." *American Economic Review,* June 1975, 492–97.

18. InterStudy. *HMO Growth: 1977 to 1978.* Excelsior, Minnesota: InterStudy, 1978.

19. ———. *The Competitive Impact of Health Maintenance Organizations: Minneapolis-St. Paul.* Excelsior, Minnesota: InterStudy, 1978.

20. Kmenta, Jan and Roy F. Gilbert, "Small Sample Properties of Alternative Estimators of Seemingly Unrelated Regressions." *Journal of the American Statistical Association,* December 1968, 1180–200.

21. Lave, Judith R. and Lester B. Lave, "Hospital Cost Functions." *American Economic Review,* June 1970, 379–85.

22. Luft, Harold S., "How Do Health Maintenance Organizations Achieve Their 'Savings'?" *New England Journal of Medicine,* June 1978, 1336–43.

23. Lewis, Charles E., Rashi Fein and David Mechanic. *A Right to Health.* New York: John Wiley and Sons, 1977.

24. Mann, Judith K., and Donald E. Yett, "The Analysis of Hospital Costs: A Review Article." *Journal of Business,* April 1968, 191–202.

25. Newhouse, Joseph P., "Toward A Theory of Non-Profit Institutions: An Economic Model of a Hospital." *American Economic Review,* March 1970, 64–74.

26. ———, "The Economics of Group Practice." *Journal of Human Resources,* Winter 1973, 27–56.

27. Pauly, Mark V., Medical Staff Characteristics and Hospital Costs." *Journal of Human Resources,* Supplement 1978, 77–111.

28. Roemer, Milton I. and William Shonick, "HMO Performance: The Recent Evidence." *Health and Society,* Summer 1973, 271–317.

29. Russell, Louise B. *Technology in Hospitals: Medical Advances and Their Diffusion.* Washington, D.C.: The Brookings Institution, 1979.

30. Uzawa, Hirofumi, "Production Functions with Constant Elasticities of Substitution." *Review of Economics and Statistics,* October 1962, 291–99.

31. Zellner, Arnold, "An Efficient Method for Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *Journal of the American Statistical Association,* June 1962, 585–612.

32. Zubkoff, Michael, Ira Raskin, and Ruth Hanft, eds. *Hospital Cost Containment.* New York: Milbank Memorial Fund, 1978.