QUALITY/CAPACITY SUBSTITUTION IN THE DELIVERY OF MENTAL HEALTH CARE

JOHN G. CROSS

Department of Economics and The Mental Health Research Institute, The University of Michigan, Ann Arbor, MI 48109, U.S.A.

Abstract—This paper stresses chronicity as an essential aspect of mental illness, and develops an analysis of the mental health services delivery problem in which chronicity plays a central role. This emphasis leads in turn to a discussion of the importance of "treatment quality" in determining long run health care costs. The paper focuses on two sources of inefficiency in the mental health services system: the possibility that there are local optima in services delivery cost functions, and the fact that separately-managed treatment centers are interdependent because the treatment practices of one affect the patient flows of others. A formal analysis of these problems is provided, supplemented by a simulation model that can provide a much more detailed description of outcomes and costs. Finally, there is a short discussion of the use of subsidization to offset the effects of interdependence.

1. INTRODUCTION

Most of the literature in medical care economics focuses on the provision of episodic medical treatment, and not on care for chronic illness. This is a severe limitation in many branches of medicine, but it is particularly restrictive in the analysis of mental health care, for which chronicity is the rule rather than the exception. Moreover, the medical care literature generally presumes the existence of more or less standardized treatment methods (possibly graduated in quality) with known consequences, or at least with known probability distributions defined over possible outcomes. In contrast, diagnosis in mental health areas is often quite uncertain, and treatments for similar disorders vary widely among institutions and among providers with different treatment philosophies.

Chronicity imposes special economic burdens on any health care system. In the first place, it limits the potential of insurance for providing private care, because the expenses of future treatments are not insurable risks. In the second place, chronicity limits the private economic capacity of those who are afflicted, particularly in the case of mental illness. Thus, while a network of private and publicly subsidized private institutions has developed to meet needs for unanticipated episodic medical care, the chronically mentally ill are treated for the most part in public institutions and at public expense. Because of this, conventional competitive market analyses are probably less relevant to mental health care delivery than are discussions of the managerial problems of providing mental health care in the presence of exogenous demand and considerable uncertainty concerning both diagnosis and treatment effectiveness.

The purpose of this paper is to focus on the problem of "quality" of care and its implications for health care costs in general. Although quality often arises as a variable in analyses of the demand for medical care (as in Feldstein [1]), the question of quality of care rarely arises in the literature on mental health. A few exceptions are found in the work of Conley [2, 3], Fein [4], May [5-7], McCaffree [8, 9], Rubin [10], and Weisbrod et al. [11], but even these provide only hints of the nature of the service-quality problem in this context.

The image of a hospital or physician providing standardized services may be appropriate for certain routine medical problems, but it is not generally accurate, and in the case of mental health care, it is a dangerous oversimplification. Depending on their circumstances, patients with similar disorders may be treated by any one or more of a variety of psychiatrists, psychologists, social workers, or even para-professional volunteers, and it is clear that the character of the treatment will depend on which of these providers is involved and in what setting. The cost of treatment is similarly sensitive to the choice and variety of providers, and, although there may be dispute even on this point, one is inclined to regard the more expensive services as the ones which deliver the higher "quality".

The problem would be simplified if it were possible to treat quality decisions independently of quantity of service decisions, but interactions between these two aspects of service are important. In the case of chronic illness, quality of care may not only reduce the necessary duration of treatment, it is also likely to influence future treatment needs, by altering the frequency of recurrences, by altering the severity of future illness episodes, or by altering future durations of treatment. The possibility arises that if high quality of care has a great enough effect on future treatment needs, it may be possible to substitute quality for capacity, perhaps even to the point of reducing the overall cost of care. Weisbrod, in particular, suggests that this might be the case for community based
treatment settings (and Feldstein [1], provides a hint of the same possibility in more conventional hospital settings). A cost reduction may occur even if a "typical" mental patient is receiving frequent services, and is never discharged from the care system. For example, a change in inpatient treatment may reduce inpatient duration even if the patient is only transferred into a residential facility earlier than would otherwise have been the case. Although the patient remains under care, the dollars devoted to treatment have been reduced because of the lower cost of residential facilities.

One might suppose that providers in the mental health care system would already have exploited opportunities to substitute quality for capacity, partly from a desire to reduce their own costs of delivering services and partly out of a desire to provide high quality services to their clients. There are several obstacles to effective quality delivery in mental health systems, however:

- Most providers have great difficulty in obtaining and assimilating the information on which the ideal quality decision should be based. Evaluation of the relationship between treatment character and overall health care needs requires statistical analysis of patient populations (within diagnostic categories) that are larger than most providers encounter. Even if adequate numbers of patients were available, the necessary statistical evaluations are unlikely. By training as well as by inclination, most mental health providers view their patients as individuals, each with individual problems and in need of individualized treatment, and this point of view often precludes the development of even elementary statistical evaluations of treatment effectiveness.
- Mental health delivery cost functions may appear to be minimized at certain quality levels even though much lower costs are obtainable at other points. A provider might actually experiment with small variations in quality to see how these changes affect costs, and if these experiments lead to increased costs, conclude that no further changes in the nature of treatment would be economically efficient. If the benefits from increased quality arise only if the increase is much larger than has been tried, day-to-day experience will not be a good guide to an optimal mix of services.
- The influence of quality over capacity is dynamic in character, and if it requires significant amounts of time for the benefits of increased quality to be realized, providers may fail to recognize them.
- The individual provider who raises service quality may not be the one who experiences the resulting reduced capacity cost. High-quality outpatient or residential care may serve primarily to reduce the clients' needs for inpatient services, for example, and thus the cost benefits of the higher quality are received by institutions other than the ones that provided it.

2. THE MENTAL HEALTH DELIVERY COST FUNCTION

A "local" optimum is a state that appears to be best (according to some criterion) because small deviations from that state all lead to inferior outcomes. If small increases or decreases in treatment expenditures fail to be worthwhile on some benefit/cost criterion, it is natural to extrapolate from this experience to the conclusion that large deviations would be even less desirable. Such a conclusion is not logically necessary, and its acceptance may lead one to overlook the possibility that large changes in expenditures (either up or down) might bring about significant improvements in treatment performance.

The possibility that experience with local optima might interfere with the achievement of a "global" optimum is one which is routinely acknowledged in decision theory, but it is not usually given extensive treatment because it often seems to involve special cases: unusual functional forms, or implausible cost conditions. In the analysis of quality/capacity substitution in mental health care, however, the possibility arises so naturally that one might expect it to be the general case. The nature of this possibility can be described with the very simple model described in this section.

In practice, one observes a variety of different lengths of stay for different patients, and a complete description of the mental health care delivery problem would require the analysis of a distribution of values for treatment duration. It is sufficient here, however, to represent this distribution with its expectation, T, which represents the mean duration of stay for patients in a particular unit. If N new patients are admitted each day (N may be less than 1), then the facility will be occupied by NT patients at any point in time. Normally, of course, the capacity of the facility would be maintained somewhat above this level in order to accommodate stochastic fluctuations in admissions and discharges. It is unnecessary to introduce this detail here, and so NT can be used as the available patient capacity of the unit.

"Quality" is a difficult concept to define because it has so many aspects. In the first place, there are several different types of input that contribute to treatment quality, so that one should consider quality to be a vector of inputs rather than a single-dimensional variable. In addition, the term is often applied to the output of mental health service as it is to the inputs. Deliberate increases in length of stay may be interpreted to be increases in quality: if a patient is retained in an inpatient setting until he/she achieves a relatively high level of functioning, then the treatment might be regarded as of higher "quality" than another that discharged the patient earlier, even if the day-to-day ingredients of the two treatments are identical.

It is sufficient for the purposes of this section to focus on inputs to quality and to standardize the output, the objects of study are then alternative ways of treating similar patients that ultimately may bring them to the same level of functioning. There are still many different kinds of input that contribute to treatment quality in this narrow sense. The elementary model used here does not require much detail, however, and can be described adequately with a hypothetical idealized single-dimensional variable Q. Lower case q will represent the price of one unit of this good per patient day. (For many purposes, it is convenient to normalize q to equal 1, so that Q is
simply dollars spent on quality). If the patients under treatment have an average capacity cost of \( P \) dollars per patient per day, then the total long run cost of operating the facility is \( PNT + qQN(T) \) dollars per day.

There is naturally some service component in any provision of capacity, and so the definition of zero quality is essentially arbitrary. Define capacity alone as the definition of minimal service quality does not change, it is possible to represent the dependence of \( T \) on values of \( Q \) through a function \( f(Q) \):

\[
T = f(Q)
\]  

where \( T_0 = f(0) \).

If the output of treatment is standardized at some specified level of functioning, then improvements in service (apart from increased lengths of stay) will tend to reduce mean durations of stay. This implies that the function \( f(Q) \) has the property:

\[
f'(Q) \leq 0.
\]  

Many mental health practitioners believe that there is some threshold of care below which mental health service is quite ineffective; below this threshold, increments in quality would have less effect than they would if the basic level of service were higher. This implies that, at least over some range, increases in quality might be increasingly effective in reducing the duration of treatment. Eventually, however, this effect must begin to disappear. It is inevitable that diminishing returns will set in at some point; otherwise it would be possible to reduce the treatment duration to zero, and of course, that would imply no treatment at all. Indeed, it is likely that durations of stay could not be reduced below some fixed lower bound no matter how many quality resources were expended. All this leads to condition (3) below:

There exists some \( Q' > 0 \) such that

\[
f''(Q) > 0 \quad \text{for all } Q > Q'.
\]  

If \( C \) is defined to be total cost per admission, then:

\[
C = (P + qQ)f(Q).
\]  

When \( Q \) varies, total cost will respond according to the derivative:

\[
\frac{dC}{dQ} = (P + qQ)f'(Q) + qf(Q).
\]  

The first term on the right-hand side of eqn (5) is negative, reflecting the fact that money is saved as treatment is reduced. The second term is positive, reflecting the cost of quality itself. Thus an increase in \( Q \) increases cost per admission in proportion to the length of time in treatment, \( f(Q) \), but the concomitant reduction in treatment days, \( f'(Q) \), reduces cost in proportion to the amount already being spent per day. The important possibility is that the cost saving component may become larger as \( Q \) is increased. When \( Q = 0 \), total cost is \( PT_0 \) and \( dC/dQ = Pf'(Q) + qT_0 \). If \( qT_0 \) is large relative to the product \( Pf'(Q) \), then \( dC/dQ > 0 \), implying that increases in quality will bring about increases in cost. Unless the absolute value of \( f'(Q) \) decreases sharply in response to increases in \( Q \), however, the product \((P + qQ)f'(Q)\) will become more strongly negative as \( Q \) becomes larger, while the absolute magnitude of \( f(Q) \) in the positive term diminishes. Eventually the sign of \( dC/dQ \) may be reversed, implying that now cost will be reduced by increases in quality. At large values of \( Q \), of course, the condition \( f'(Q) > 0 \) will reduce the magnitude of \( f'(Q) \) by so much that the sign of \( dC/dQ \) is reversed again, but before this has happened, total cost may well have fallen below \( PT_0 \), which is the cost incurred when \( Q = 0 \).

Unless the cost minimum does occur at \( Q = 0 \), the least-cost supply of quality, \( Q^* \), must satisfy the first and second order conditions (6) and (7) respectively:

\[
(P + qQ^*)f'(Q^*) + qf(Q^*) = 0,
\]  

\[
2qf'(Q^*) + (P + qQ^*)f''(Q^*) > 0.
\]  

A general review of the types of cost function that can occur is provided by Fig. 1. One can distinguish four general forms (although more elaborate mixtures of these are, of course, possible):

I. Initially Declining. This is a uniformly convex function with the property that costs decline monotonically with increases in quality expenditure until the unique local minimum is achieved.

II. Initially Rising, then Strongly Falling. This is a function that rises at first, but then falls enough for cost levels to be achieved that are below those at the 0-quality level.

III. Initially Rising, then Weakly Falling. This function is similar to II, except that costs do not fall back to the 0-quality level.

IV. Monotonically Rising. This, like Case I, is a uniformly convex function, but in this case, costs rise with all increases in quality.

The distinction between a local and a global minimum in Cases II and III arises because the cost function described by eqn (4) is not necessarily convex. The nonconvexity is due to the product form of the cost function. The existence of multiple optima would assume less importance were the nature of the function \( f(Q) \) well known in advance. If providers and provider managers were informed as to the precise relationship between treatment quality and treatment duration, they would have no difficulty in specifying the cost function and calculating the cost-minimizing level of quality for themselves. Unfortunately, not only is this information unavailable, treatment effectiveness is a matter of considerable

![Fig. 1. Possible forms of the cost function.](image-url)
controversy, and at the present level of knowledge, providers are not able to agree on even rough approximations to the values of \( f(Q) \) (beyond acknowledgement that \( f'(Q) \) is likely to be negative). As a consequence, practical experience is the only available guide to effective service delivery, and since few would be willing to experiment with wide variations in quality expenditure, the lure of the local cost minimum would be hard to resist.

3. EXTERNAL BENEFITS

A second major impediment to efficient quality delivery stems from the fact that the cost reductions attributable to increased quality may not be received by the agency that makes the quality expenditure. If high quality outpatient care reduces the need for inpatient services, for example, decreases in overall mental health care costs may be associated with increases in outpatient costs, and the interests of those who pay for the health care system (through insurance or public funds) would not be consistent with the interests of outpatient providers.

To the extent that individual providers do seek to restrain costs, there is an unambiguous bias toward the underprovision of quality from a system point of view. Suppose that patients are admitted to and discharged from each treatment center according to established and unchanging criteria for each level of impairment. Then the patient “demand” on any single provider or health care agency will depend on the quality choices of all providers—low quality offered by one agency will increase the need for services provided by others, while high quality will reduce the demands placed on others. This proposition can be expressed formally: letting \( Q_j \) represent the choice by agency \( j \) of quality per patient-day, the demands placed on an agency \( j \) will be a function of the vector of quality choices made by all \( n \) agencies:

\[
D_j = f(Q_j), \quad \text{where} \quad Q_j = Q_1, \ldots, Q_n. \tag{8}
\]

By the nature of the situation, one must retain counterparts to conditions (2) and (3):

\[
\frac{\partial D_j}{\partial Q_i} = f_i(Q) \leq 0 \quad \text{for all } j. \tag{9}
\]

There exists \( Q^* = Q_1, \ldots, Q_n \) such that:

\[
\frac{\partial^2 D_j}{\partial Q_i^2} = f_{ii}(Q) > 0 \quad \text{for all } Q > Q^*. \tag{10}
\]

If each agency \( j \) faces a capacity cost given by \( P_j \), then system costs are:

\[
C = \sum_{j=1}^{n} (P_j + qQ_j)f_j(Q), \tag{11}
\]

and overall mental health care costs are minimized if the \( n \) quality choices satisfy:

\[
\frac{dC}{dQ_i} = (P_i + qQ_i)f_i(Q) + \sum_{j \neq i} (P_j + qQ_j)f_j(Q) + qf_i(Q) = 0. \tag{12}
\]

for all \( i = 1, \ldots, n \).

The costs borne by agency \( i \) alone are given by \( C_i = (P_i + qQ_i)f_i(Q) \), and if this provider responds only to these local costs, \( Q_i \) will be chosen so as to minimize \( C_i \). That is, \( Q_i \) will satisfy:

\[
\frac{dC_i}{dQ_i} = (P_i + qQ_i)f_i(Q) + qf_i(Q) = 0. \tag{13}
\]

If the strict inequality in eqn (9) holds, then it follows that satisfaction of eqn (13) will not lead to the satisfaction of eqn (12), but to the condition \( dC_i/dQ_i < 0 \) instead. Thus each agent \( i \) will underprovide quality from the system point of view.

4. THE PERSPECTIVE FROM SIMULATION

The demonstrations of the possibility of inefficient quality delivery that have been given above have rested on general models that capture plausible aspects of the mental health delivery process. Nevertheless, these models do not have a firm quantitative base. The state of knowledge concerning the effectiveness of mental health treatment is not nearly advanced enough to permit specification of the analytical form of the mental health cost function, or to characterize the long-term consequences of intensive, high-quality treatment.

There is an alternative approach to the analysis of this problem, however, that can take one much closer to empirical reality. It is possible to represent the mental health delivery system by means of computer simulation models, and to test the consequences of changes in quality delivery through changes in the model parameters. This approach still requires simplifications in the definition of “quality” and in the representation of different treatment types, but it does afford a more detailed analysis of patient courses through the health care system than is available through functional analysis alone.

At the Mental Health Research Institute at the University of Michigan, we have developed a number of stochastic models of community mental health services that permit the introduction of a great deal of empirical realism. These models provide for a variety of different types of treatment centers, and permit evaluation of the interdependencies among these centers. They also distinguish among different levels of severity or stage of illness within each treatment center. At the core of these models is a matrix of Markov transition probabilities that describe the likelihoods with which patients change illness severity or treatment location. These probabilities have been inferred both from actual patient movements within community mental health systems in Michigan and from the judgments of clinicians and evaluators with experience in the field. These probabilities can be made to vary with changes in per-patient treatment expenditures, and the resulting system can be used as a simulation of what might happen if these changes in treatment expenditure were actually made.

For the purposes of this paper, a simple simulation model was developed using as its providers a pair of outpatient centers, two community residential facilities, and two different types of inpatient setting. The object of the simulation was to investigate the impact on overall system cost of increased expenditure on treatment quality in one of the outpatient facilities.
Specific questions raised by the analysis in this paper are:

1. Is it possible for increases in treatment quality expenditure to reduce mental health care costs?
2. If increases in treatment quality expenditure reduce costs, are the cost reductions received by the treatment center that delivers the treatment, or are these benefits primarily received elsewhere?
3. Do increases in treatment quality tend to increase system costs at low levels but decrease costs at higher levels?

The simulations suggest affirmative answers to all three of these questions, but they suggest that by far the most important impediment to efficient quality delivery is embodied in the second: the fact that the benefits from improved treatment quality are received by centers other than the one providing enhanced services. Figures 2 and 3 below describe typical simulation outcomes. On the horizontal axes of the figures is a scale factor that indexes changes in treatment quality in one of the two outpatient settings. A value of 1.0 corresponds to the estimate of current practice. On the vertical axis of Fig. 2 is the total cost of that outpatient center: not surprisingly, this cost is almost perfectly proportional to the level of treatment quality. On the vertical axis of Fig. 3 are total system costs—including the expenses of the high-quality outpatient center that it contains. These costs are by no means proportional to the level of treatment quality in the outpatient center, and in fact they are strongly decreasing over most of the range of the simulation. Moreover, the decrease extends well beyond the point corresponding to current practice. It is clear, however, that those who manage the outpatient center have no incentive to deliver these increases in quality: the savings in system costs are generated only from increases in outpatient expenditures. At the simulated value of the status quo, every dollar spent on improved outpatient treatment reduces system costs by 2 dollars (and since the outpatient cost is included in the system cost, that means that every dollar spent in the outpatient setting saves 3 dollars elsewhere). Unless some means were found to shift these savings to the center that generated them, no further improvement in treatment quality is likely to take place.

The simulations also display the nonconvexity suggested by eqn (4), but the range of nonconvexity corresponds to quality levels that are lower than those that we have estimated to be available in practice—system costs do rise with increases in quality at low levels, but the range is too small to show clearly on the scale of Fig. 3. This may, of course, be a reflection of optimism on the part of those who assisted in the establishment of the parameters of the model. In any case, this stands as a preliminary exercise, and its value is primarily that of showing that the nonconvexity is present in models that possess a good deal of empirical realism.

### 5. SUBSIDIZATION FOR EFFICIENCY

The simulations suggest that various types of mental health service units are strongly interdependent because each influences the patient flows to each of the others. To the extent that the managements of these facilities focus on problems of cost and services delivery within their own units, this interdependence will be ignored, and the overall cost-effectiveness of the system will be severely impaired. The problem is not one of information: experienced managers are surely aware that their decisions affect the patient loads in other facilities. The problem is that each manager’s primary responsibility is the successful operation of one unit, and any system that evaluates the performance of a manager will reflect that fact. Even forceful demonstrations of the importance of the interdependence problem will not reorient the incentives that are leading to the inefficiency.

One solution might be the integration of the entire mental health services system under one manager/director. Taking full account of the interdependence among facilities would then become part of the responsibility of this single individual. Unfortunately, this approach does not solve the problem so much as it redefines it; how is the single director to create incentives in the system so as to induce unit managers (who must still exist) to operate as though they also cared about the interdependencies? In practice, this is evidently very difficult, in that community mental health systems with managing directors still operate more or less as collections of autonomous service units.

One interesting way to address this problem is to introduce a standard economic device: within the context of traditional market economies, it is often
suggested that external interdependencies of this sort may be corrected by instituting systems of taxes and subsidies that reflect the external costs or benefits. Such a scheme would entail the creation of a "quality subsidy" where each treatment center is paid an extra sum that is proportional to the rate of quality delivery. The funding agency would invent a subsidy rate that approximated as closely as possible a value $S_i$ for each provider $i$, where

$$S_i = - \sum_{j \neq i} (P_j + qQ_j)f_j(Q_j). \quad (14)$$

The subsidy $S_i$ is equal to the aggregated cost savings to agencies other than $i$ that are realized when $i$ increases its quality per patient day by one unit. If agency $i$ is granted this subsidy, then its cost function becomes:

$$C_i = (P_i + qQ_i)f_i(Q_i) - S_iQ_i, \quad (15)$$

and it is easy to see that if such a subsidy is granted to each treatment center, and if each follows a cost minimizing quality choice, then these decisions together will satisfy the efficiency condition in eqn (12).

Subsidization of mental health services by various government agencies and programs is common practice, but to the author's knowledge, none of them attempt to approximate $S_i$ as it is defined in eqn (14). Instead, they are inclined to subsidize the inputs to quality, such as staff or accommodations. This state of affairs is easy to understand: quality per patient-day is much more difficult to measure objectively than are expenditures on staff or facilities, and a correct estimation of $S_i$ requires an analysis of mental health systems that is sufficient to support quantification of the interdependence problem.

Understandable as this choice of subsidies may be, it is not at all efficient. Indeed, the point of the argument here is that the provision of quality per patient-day is not the same as total resources devoted to quality, and that the two may actually move in opposite directions. Inputs to quality do not vary in proportion to $Q_i$, but in proportion to $Q_if_i(Q_i)$. Indeed, as $Q_i$ is increased, the mean duration of stay, $f_i(Q_i)$ decreases (because of the negative slope of the function $f_i$), and therefore the product $Q_if_i(Q_i)$ increases less rapidly than does $Q_i$ alone. In short, a subsidy on the inputs to quality will encourage providers to deliver less quality than they would if the subsidy were directed at $Q_i$ alone.

A further pitfall to providing subsidies to the inputs to quality is that there is no assurance that they will be used efficiently. In effect, subsidies on inputs reward treatment centers for displacing other resources with the subsidized inputs in a way that minimizes the improvement in treatment quality (that is, in a way that minimizes the reduction in treatment duration).

Finally, no subsidy system can work without facility managers with strong incentives toward cost minimization. This requirement may be severely compromised in the present-day climate of medical cost reimbursement plans. If costs are recovered in any case, there is no cost-reduction incentive system in force, and a managerial preference for large staff and facilities will conflict with efficient service delivery.

Moreover, under many existing financing schemes, the subsidy would simply replace a portion of some cost-covering reimbursement, and the incentive structure faced by agency managers would be unaffected.

REFERENCES

11. B. A. Weisbrod, M. A. Test and L. I. Stein. Alternative

