GLOBAL BUDGETS AND EXCESS DEMAND FOR HOSPITAL CARE

ROGER FELDMAN[1]* AND FELIX LOBO[2]
[1]*University of Minnesota, Minneapolis, MN, USA*
[2]*Universidad Carlos III, Madrid, Spain*

SUMMARY

Excess demand is a pervasive feature of health care systems that use global budgets to pay for hospital care, regardless of the amount of money spent by those systems. This paper presents a theory that explains this feature of global budgets. The theory emphasizes that hospital administrators control the allocation of their budget, and that they choose quantity and resource intensity to maximize their own utility. The equilibrium quantity of care provided may be less than quantity demanded by consumers, leading to excess demand for admissions. An increase in the hospital's budget may even be associated with an increase in excess demand. © 1997 by John Wiley & Sons, Ltd.

KEY WORDS — global budgets; hospital care

INTRODUCTION

Many countries use global budgets to pay for hospital care. Canada, Spain and the United Kingdom (prior to 1991) are examples. Most global budget systems share two features. The first feature is a central authority which determines the budget for each hospital during a stated period of time, such as 1 year. The central authority could be the national government or, as in the case of Canada, 10 provincial and two territorial public funding systems. Particular programmes within the health care sector may also use global hospital budgets. For example, in the United States, the Veterans Administration provides hospital care for military veterans through the use of global budgets. Because it determines the budget for every hospital in the system, the central authority controls the total amount spent on hospital services, as well as the allocation of funds among hospitals and regions of the country. This is in contrast to 'open-ended' reimbursement systems such as those that utilize cost reimbursement or fixed payment per admission.

The second feature of global budget systems is that the hospital has a considerable amount of discretion over how its budget is spent within each period. Because the hospital itself exercises this control, the central authority can absolve itself from the responsibility of 'micro-managing' the production of patient care. This job is turned over to professionals who, it is assumed, are better qualified to determine the optimal quantity and quality of care and to decide which patients deserve to receive care. Typically, price plays no role in this process because hospital care is free

*Correspondence to: Dr Roger Feldman, Division of Health Services Research, University of Minnesota, 420 Delaware Street, SE, Box 729, Minneapolis, MN 55455, USA. Tel. (612) 624-5669; Fax (612) 624-2196; e-mail feldm002@maroon.tc.umn.edu.

for all patients covered by the health care systems of the countries mentioned above.

Although these arguments appear to provide a powerful justification for global budgets, the experience of countries utilizing this payment system is one of excess demand for hospital services. For example, Naylor[1] reported that 1700 per on were waiting for open heart urgery in Ontario, Canada, in 1989. This represented more than 25% of the annual provincial caseload for open-heart surgery. Waiting times for elective cases ranged from as little as 4 to 8 weeks in some centres to 6 months or more in others. Other comparisons indicate that the waiting time for knee replacement is longer in Ontario than in the United States,[2] and the waiting time for hip fracture surgery is longer in Quebec than in Massachusetts.[3]

Frankel[4] summarized the official data on inpatient waiting lists in the United Kingdom's National Health Service. As of 31 March 1991, about 690 000 people were waiting for some type of inpatient medical admission. According to Department of Health figures,[5] 169 800 of these people (almost 25% of the total waiting list) had waited over 1 year for treatment. Of interest to the hypothesis advanced in this paper, time series data indicate that the NHS hospital waiting list has remained remarkably constant since about 1960, when measured as a percentage of total hospital throughput. This has occurred despite swings in real funding for the NHS, suggesting that hospital waiting lists are a persistent characteristic of the NHS, irrespective of the level of resources devoted to hospital care.

Why is excess demand a pervasive feature of globally budgeted hospital payment systems? Pauly[6] has suggested that the problem is a generic one that applies to any public program of 'free' medical care. The public, as taxpayers, are not willing to spend as much money as they demand in their role as consumers of services. According to Pauly, the result is chronic under-funding of free health services. Buchanan[7] made a similar argument which relies on altruism (or the lack of it) to explain under-funding of public services. As taxpayers, consumers are asked to vote for a tax contribution that largely pays for the care of others. Unless individuals consider the consumption of others to be equally important as their own consumption, they will not vote to supply as many hospital services as they demand. Buchanan's argument, unlike that of Pauly, does not depend on the health services being 'free' on demand. However, if the services are free, then the case for under-funding is strengthened because people will not value even their own marginal consumption at its marginal tax cost.

In this paper, we present a theory that explains why hospitals appear to be chronically under-funded in systems as different as the United Kingdom, which spent 6.2% of its Gross Domestic Product (GDP) on health care in 1990 and Canada, which spent 9.3% of GDP on health care.[8] Surprisingly, our theory suggests that an increase in the amount of money budgeted to a hospital may even *increase* the excess demand for hospital services. This is in contrast to the implication of Pauly's and Buchanan's models, in which chronic under-funding can be corrected, at least in principle, by infusing money into the system.

BASIC MODEL OF A GLOBAL BUDGET FOR A HOSPITAL

Equilibrium in the basic model

In this section we outline a simple economic model of a hospital in a global budgeting system, and we show that excess demand may be an equilibrium of this model. The concept of equilibrium refers to a situation in which certain inter-related variables in a model are selected so that no tendency to change prevails.[9] Our model focuses on two inter-related variables: the quantity and quality of hospital services. We assume that hospital administrators select these variables with the goal of maximizing their own utility, subject to patients' demand and the size of the hospital's budget. Thus, equilibrium in our model refers to a combination of quantity and quality which, once chosen, will not change as long as the external forces facing the hospital are assumed to be fixed. However, we will show that this equilibrium may not coincide with the quantity and quality of hospital care that consumers would prefer for the same global budget.

Our model begins with the assumption that demand for services at a particular hospital (Q_d) depends on the patients' price of care at that hospital (P) and their perceptions of its quality: $Q_d = D(P,q_c)$, where q_c stands for quality as perceived by patients. However, in the health systems analysed in this paper (e.g. Canada and the UK), the patients' price is zero. Consequently, we hypothesize that the demand for services at a particular hospital depends entirely on perceived quality.

Although perceived quality of hospital care is not observed directly, we suggest that it is positively related to the intensity of resources utilized per unit of service (e.g. per admission or patient-day). The notion that resource intensity can be used as a proxy for hospital quality was popularized by Feldstein,[10] who proposed to measure quality as the ratio of an index of average cost per patient-day to an index of resource prices. Feldman and Dowd[11] measured quality by the number of specialized facilities and services and whether the hospital had a teaching programme. It was found that offering more services always increased the demand for admissions, especially by Medicare patients. Since Medicare patients pay a fixed deductible per hospital stay, they should exhibit quality-sensitive demand when choosing among hospitals (i.e. they should prefer the hospital with the highest perceived quality).

The role of perceived quality in explaining consumers' hospital choices has attracted the attention of market researchers,[12–15] who consistently find that perceived quality is the most important attribute in hospital selection. One study[15] showed that perceived hospital quality could be measured by a valid and reliable scale composed of 16 indicators. These included, in addition to a direct perception of 'quality of medical care', items such as the level of technology, range of services and hospital size. These measures all relate to resource intensity.

A final perspective on the relation between quality and resource intensity has been offered by Phelps,[16] who notes that doctors are an important input in the production of medical care. A primary way by which hospitals attract doctors to their staffs is to provide the capability for doctors to do things they cannot do elsewhere. This may include the provision of a cardiac intensive care unit to attract cardiologists, for example. Doctors then steer patients toward hospitals that offer these resource-intensive services. The implication of this perspective is that doctors can also use resource intensity as a proxy for hospital quality even in situations where patients are not well informed.

On the basis of these studies, we write the patients' demand function as $Q_d = F(R)$, where R is the intensity of resource use, and we assume that $F'(R)$ is positive. Nonetheless, after developing the basic model, we analyse the case where patients perceive that the hospital is providing excessive services [i.e. the demand curve is backward-bending so that $F'(R)$ is negative].

Next, we explain our assumption regarding hospital decision-makers' preferences. We assume that the hospital maximizes a utility function that depends on quantity of services and quality, as perceived by the professionals who run the hospital: $U = H(Q,q_h)$, where q_h stands for professional perceptions of quality. These perceptions are not observed directly and may differ from consumers' perceptions, but like consumers' perceptions, we assume they are positively related to observed resource intensity. Because resource intensity is a proxy for q_h, utility can be mapped into a new function, $U = V(Q,R)$, where q_h has been replaced by R.

To find the equilibrium of our model, we want to use indifference curves from this new utility function because it can be drawn in a graph with the same dimensions (quantity and resource intensity) as consumer demand. However, indifference curves from the V function may not be convex to the origin, in which case the usual tangency condition for equilibrium would not apply to our model. In the Appendix, we derive a twofold sufficient condition for convexity of indifference curves from V: (1) indifference curves from the underlying utility function, $H(Q,q_h)$, are convex; and (2) increasing the level of resource intensity results in a proportionately equal or lesser increase in professional perceptions of quality (i.e. the relation between resource intensity and perceived quality displays constant or decreasing returns). Since these are reasonable assumptions, we will draw the indifference curves from V as convex to the origin of a graph whose dimensions are quantity and resource intensity.

The next element of our model is the hospital's budget constraint. In a global budget system, the hospital's budget is $B = RQ$, where B is assumed to be determined exogenously by the central authority. After presenting the basic model, we will consider refinements in which future budgets depend, at least in part, on current actions taken by the hospital.

The hospital is assumed to maximize utility, subject to the budget constraint. This is a standard Kuhn–Tucker problem, which can be represented as maximizing the Lagrangian function:

$$L = V(Q,R) + \lambda(B - RQ) + \gamma[F(R)-Q] \quad (1)$$

The first-order conditions for utility maximization are given by

$$\frac{\partial L}{\partial Q} = V_Q - \lambda R - \gamma \leq 0,\ Q \geq 0,\ Q\frac{\partial L}{\partial Q} = 0 \quad (2a)$$

$$\frac{\partial L}{\partial R} = V_R - \lambda Q - \gamma F_R \leq 0,\ R \geq 0,\ R\frac{\partial L}{\partial R} = 0 \quad (2b)$$

$$\frac{\partial L}{\partial \lambda} = B - RQ \geq 0,\ \lambda \geq 0,\ \lambda\frac{\partial L}{\partial \lambda} = 0 \quad (2c)$$

$$\frac{\partial L}{\partial \gamma} = F(R) - Q \geq 0,\ \gamma \geq 0,\ \gamma\frac{\partial L}{\partial \gamma} = 0 \quad (2d)$$

We will assume that both decision variables (Q and R) are strictly positive, which implies that the respective partial derivatives of the Langrangian function (2a and 2b) are zero. This equilibrium for the model is shown in Fig. 1. Assuming that Q and R are strictly positive implies that the utility function $V(Q,R)$ is more convex that the budget constraint, which is indicated by the rectangular hyperbola B_0 in Fig. 1.

Next, consider the conditions governing λ and γ:

(i) *Suppose* $\lambda = 0$*:* then γ must be positive to make equation (2a) equal zero, but γ must be negative to make equation (2b) equal zero. This is a contradiction, so $\lambda > 0$. This implies, by equation (2c), that the hospital's budget constraint is binding.

(ii) *Suppose* $\gamma = 0$*:* in this case we can solve equations (2a) and (2b) to obtain $V_Q/V_R = R/Q$. Condition (2d) implies that $F(R) - Q \geq 0$.

The assumption that $\gamma = 0$ means that patients' demand is not binding; therefore, hospital administrators can choose quantity and quality as they wish, subject only to the size of their budget. Equilibrium occurs where an indifference curve from V, labelled V_0, is tangent to the budget constraint. Since this occurs to the left of the intersection of the budget constraint and the demand curve, there is excess demand for serv-

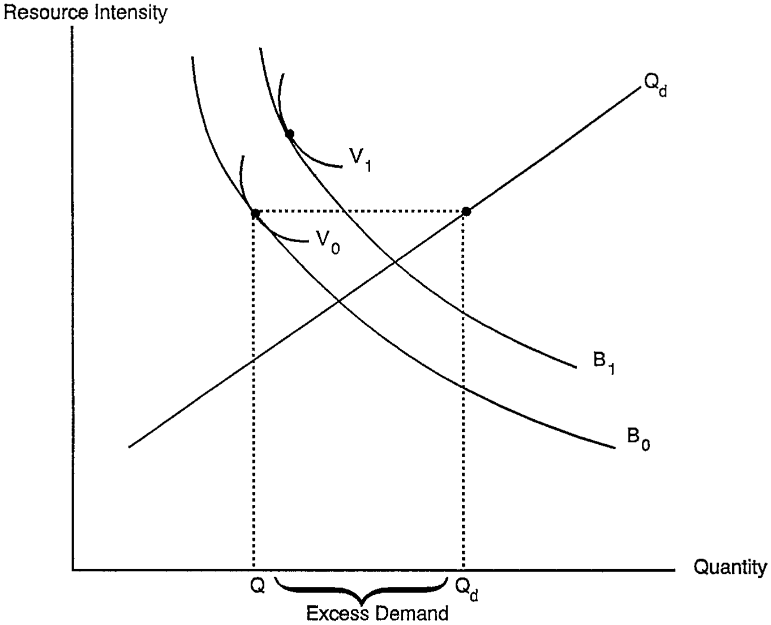

Figure 1. Excess demand for hospital services in a global budget payment system.

ices, measured by the horizontal distance ($Q_d - Q$) in Fig. 1.

A unique feature of this model is that the demand curve Q_d is upward-sloping. This appears to violate the economic law that demand curves for non-Giffen goods slope downward. However, it is important to remember that the vertical axis of Fig. 1 represents resource intensity (a proxy measure of quality), not the price of hospital care. The demand for hospital admissions should increase as the quality of services increases. In the usual graph of demand as a relation between price and quantity, the demand curve would appear to shift to the right when quality increases.

It is possible that $\gamma = 0$ simultaneously with $F(R) = Q$ (i.e. the patient demand constraint is not binding but there is no excess demand). Graphically, this would be a corner solution where V_0 is tangent to the budget constraint precisely where Q_d crosses the budget constraint in Fig. 1.

(iii) *Suppose $\gamma > 0$:* in this case condition (2d) implies that $F(R) - Q = 0$ (i.e. there is no excess demand). Conditions (2a) and (2b) can be solved to obtain

$$\frac{V_Q}{V_R} = \frac{R}{Q}\left(1 + \frac{\gamma F_R}{V_R} + \frac{\gamma Q}{RV_R}\right) > \frac{R}{Q}$$

The assumption that $\gamma > 0$ means that patients' demand is binding. This equilibrium can be interpreted with reference to the original situation shown in Fig. 1. Imagine that a different hospital, facing the same demand curve and the same global budget, wants to stretch that budget toward more quantity and less quality. Tangency between one of the hospital's indifference curves and B_0 may occur to the right of the intersection between B_0 and Q_d. However, patients would find this level of quality unacceptably low, so they would not demand as much care as the hospital wants to produce. The result would be underutilization of the hospital's budget and less utility than the hospital could achieve by increasing the quality of care. The best possible combination of quality and quantity for the hospital under these new assumptions would occur at the intersection of B_0 and Q_d.

There is a wide variety of utility functions that will lead to an equilibrium of the type just described (a corner solution at $Q = Q_d$, with both constraints binding, and no tangency at the equilibrium). This implies that excess demand is not always the equilibrium solution to our model. Nevertheless, we believe that administrators and patients would be at odds in this corner solution. The administrators would like to provide high volume and low quality, but patients would not willingly utilize this amount of low-quality care.

Hospital response to an exogenous budget increase

As explained above, our model predicts that equilibrium with excess demand will always occur when the hospital's marginal rate of substitution is strictly less than the slope of the budget constraint at the point where $F(R)$ crosses the budget constraint. In this section, we show that excess demand may increase following an exogenous increase in the hospital's budget, and we derive the condition under which this 'perverse' response will occur.

By definition, excess demand is ($Q_d - Q$), so excess demand will increase as B increases if

$$\frac{\partial(Q_d - Q)}{\partial B} = \frac{\partial Q_d}{\partial R}\frac{\partial R}{\partial B} - \frac{\partial Q_d}{\partial B} > 0 \tag{3}$$

Next, totally differentiating the budget constraint:

$$1 = R\frac{\partial Q}{\partial B} + Q\frac{\partial R}{\partial B} \tag{4}$$

Substituting equation (4) into equation (3), we can rewrite the resulting formula in elasticity notion as

$$\varepsilon < \frac{\eta}{(Q/Q_d) + \eta} \tag{5}$$

where ε is the hospital's elasticity of supply of services with respect to the budget and η is the consumers' elasticity of demand for services with respect to resource intensity.

Suppose that the initial equilibrium involves a perfect balance between supply and demand (i.e. $Q/Q_d = 1.0$). Then, any supply response less than $\eta/(1 + \eta)$ will be insufficient to satisfy the increased demand for services. For example, if

Q/Q_d = 1.0 and η = 1.0, then an increase in the hospital's budget will result in more excess demand if the elasticity of supply is less than 0.5. In the extreme case where demand is infinitely elastic, excess demand will increase if the elasticity of supply is less than 1.0. In other words, when demand is infinitely responsive to quality, the increase in the hospital's budget must be spent entirely on more services or else excess demand will increase.

Now, suppose we start from an initial position where excess demand is already present. In this case, the condition for preventing an increase in the hospital's budget from worsening the excess demand situation is even more stringent. For example, suppose that the initial position is Q/Q_d = 0.5 and η = 1.0. Then, any supply elasticity less than 0.67 will cause excess demand to increase. The intuition behind this result is that an increase in demand will out-pace a proportionate increase in supply when the initial level of demand exceeds supply. Thus, supply must expand by a larger proportion in order to 'stay even' with the increase in demand.

An alternative explanation of equation (5) is that excess demand increases with increases in the budget when the hospital's income expansion path is steeper than the $F(R)$ function. For example, $\partial(Q_d - Q)/\partial B$ is greater than zero if Q is an inferior good to the hospital and is less than zero if R is an inferior good to the hospital. Finally, if both goods are normal, excess demand could increase or decrease with an increase in the budget depending on the relative strength of the two income effects and the slope of the $F(R)$ function.

Our comparative statics analysis implies that persistent excess demand is a pervasive feature of health care systems that use global budgets to pay for hospital care, regardless of the amount of money spent by those systems. This outcome is shown in Fig. 1 by the budget constraint labeled B_1. Although B_1 represents a larger budget than B_0, hospital decision-makers have used the additional funds to finance a substantial increase in resource intensity, which has caused excess demand to increase. The income expansion path from V_0 to the new equilibrium on V_1 is clearly steeper than the slope of the $F(R)$ function in this example.

EXTENSIONS OF THE BASIC MODEL

Endogenous hospital budgets

Our assumption that hospitals' budgets are determined exogenously may be unrealistic. In practice, the central authority may monitor some indicators of hospital performance in the current period, and it may base the budget for next period, in part, on these indicators. Hospitals in the Spanish National Health System, for example, are budgeted on the basis of standardized patient days known by the acronym UPA.[17] The hospital submits its proposed UPAs for the next year to the Ministry of Health. Its budget for the next year is based on a negotiated average of the proposed UPAs and its actual UPAs for the current year, times a standard rate per UPA. In practice, the budget tends to be very close to the current UPA workload times the standard rate.

We can represent this endogenous budget-setting process by the equation $B_{t+1} = Q_tS$, where the subscripts stand for time periods and S (without a subscript) is the standard rate set by the central authority. The central authority might also update S. For example, S could be increased by a specific percentage of the economy's inflation rate. For simplicity, we assume that S is constant.

The hospital is now subjected to a multi-period optimization problem, since the budget for next year depends on its allocation of the current budget between output and resource intensity. A typical equilibrium condition for this optimization is

$$\frac{V_{Q,t}}{V_{S,t}} = \frac{S}{Q_t} - \delta\left(\frac{V_{S,t+1}}{V_{S,t}}\right)\left(\frac{S}{Q_{t+1}}\right) \qquad (6)$$

The left-hand side of equation (6) is the slope of a hospital indifference curve in the current time period. The first term on the right-hand side is the slope of the current-period budget constraint. The second right-hand side term has three parts: δ (assumed to be ≤ 1) is the discount rate used by the hospital, $V_{S,t+1}/V_{S,t}$ is the ratio of the marginal utilities of resource intensity in the current period and the next period and S/Q_{t+1} is the marginal future return from reducing resource intensity in the current period. Multiplied together, these three parts represent the discounted rate of return (i.e. the reward in terms of future budget increases) for reducing current resource intensity.

Naylor[1] used the phrase ‘fiscal envelope’ to describe the hospital’s global budget constraint. This is an apt metaphor for the hospital with an endogenous hospital. Although the hospital could push its current resource intensity to the edge of the envelope, it has an incentive not to do so because this would adversely affect next year’s budget. The prospect of future budget increases therefore provides an incentive for the hospital to hold resource intensity below the level it would choose in a single-period model.

To continue with the example of the Spanish hospital system, if the hospital serves more than the planned number of patients, or the cost per UPA is higher than anticipated, in principle the hospital does not receive additional resources. However, in practice, the hospital always tries to negotiate additional money if it needs it. This type of ‘after-the-fact’ adjustment may cause the hospital’s actual budget to differ somewhat from the pure fiscal envelope suggested by our model.

Backward-bending patient demand

We have assumed that patients view more resource intensity as being associated with higher quality. Is this assumption pivotal in driving the excess demand results? To answer this question, we extend the model to the case where patient demand is backward-bending (i.e. increases in resource intensity are perceived as reducing quality on the margin). The first-order conditions from our basic model [equations (2a)–(2d)] still apply in this case. However, the interpretation of these conditions becomes more complex. In the basic model, the hospital’s budget constraint always was binding ($\lambda > 0$), and excess demand was the expected result. In the extended model, however, it is no longer contradictory to suppose that $\lambda = 0$, as shown by the following analysis:

Suppose $\lambda = 0$: conditions (2a) and (2b) both imply that $\gamma > 0$. These conditions can be solved to obtain $V_Q/V_R = R/Q$. Condition (2d) implies that consumer demand is binding.

This equilibrium is shown in Fig. 2, where the backward-bending demand curve Q_d cuts the hospital’s budget at two points labeled α and β. The demand curve is tangent to a hospital indifference curve in the region above and to the left of β. In other words, the consumer demand constraint is binding but the hospital does not utilize its entire budget. This equilibrium would be associated with excess supply of hospital care: at the end of the budget period, hospital administrators would refund part of their budget to the central health authority. While the intuition that such refunds should not occur is fairly strong, National Health Systems managers are becoming very sophisticated and this behaviour might be possible in the future.

LINK BETWEEN EXCESS DEMAND AND WAITING LISTS

In a market, excess demand is cleared through a rise in price. However, the systems which we are analysing provide hospital care free for all patients covered. Pauly[6] maintained that the absence of prices in medical care markets leads to a generic problem of excess demand. Buchanan[7] blamed the chronic NHS funding deficiency on the absence of prices, maintaining that patient demand will persistently outpace supply in this situation. We concur with this analysis but we want to go further than Pauly and Buchanan. With zero prices and scarce resources, non-market mechanisms to allocate resources have to be implemented. Specifically, we want to discuss the relation and the consequences of two such mechanisms that happen to be implemented together in some countries: global budgets *and* waiting lists.

In more general terms, global budgets are one form of centralized decision taken by the authorities as opposed to decentralized decisions arrived at by the price system. Global budgets are not a necessary consequence of universal coverage — other ways to finance hospitals are common. In this very general setting even the endogenous Spanish budgeting discussed in the paper is a form of centralized decision making, as indicators are defined and decided centrally, although the participation of individual hospitals is obviously greater.

At the micro level, hospitals may allocate their global budget in different ways. Queuing and waiting, with patients classified by seriousness and urgency and with throughput and time as the main variables, is not the only one. Instead of waiting lists on a first-come, first-served basis, hospitals could use lotteries to match supply and demand. More realistically, they may turn to some kind of centralized decision making, but now at the hospital level, that is, giving doctors the power to select among patients with the same condition and equal urgency, which ones are to be treated first. One such criterion may be age, with younger patients preferred first; another may be the personal interests of the doctors, as may be the case when they give preference to an important politician or to the relatives of their colleagues. No doubt a mixture of these 'methods' is to be found in reality. We give preference to waiting lists.

Strictly, the existence of a waiting list is not sufficient to conclude that excess demand exists. For example, people wait for new cars to be delivered and there are waiting lists for tickets to entertainment events. These lists could represent producers' desires to avoid holding an excess inventory, or consumers' desires to plan ahead in order to consume the good at a particular time. However, given the length of waiting lists associated with some medical services, it is unlikely that waiting reflects rational planning to avoid excess capacity or to ensure consumption at a certain time. For example, half of the patients who underwent coronary artery bypass graft surgery (CABGS) in Toronto in 1989 had waiting times above the maximum indicated by medical urgency criteria.[18] Several widely publicized deaths occurred among patients who had suffered multiple waits and long cancellations.[1]

This evidence clearly indicates a link between excess demand and waiting lists in the countries mentioned in this paper. When demand is greater than supply, at least some patients are likely to be added to a list. However, the link between excess demand and the dynamics of waiting lists is less clear from our model. To illustrate this point, the model does not predict how long the waiting line will be for a given level of excess demand. This requires additional specification of the preferences of consumers and providers. Lindsay and Feigenbaum,[19] for example, pointed out that the waiting list will grow until the value of future consumption is equal to the cost of entering the

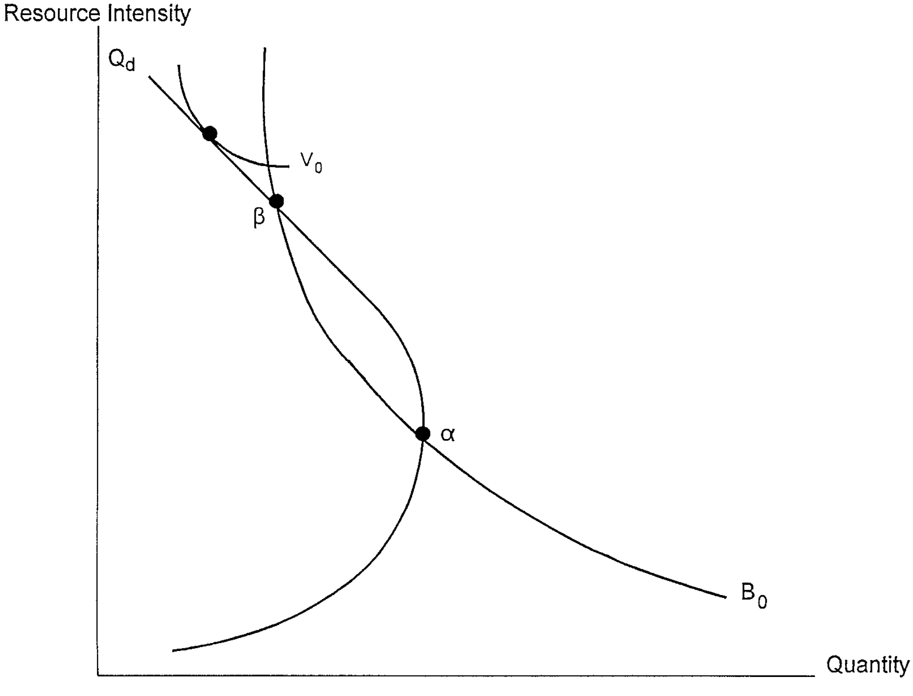

Figure 2. Equilibrium when patient demand is backward-bending.

queue for the last person to enter. Systemic differences in preferences for current use in relation to future use must be taken into account in explaining the length of the hospital waiting list. At a more detailed level, waiting times might be measured separately for patients with different acuity levels. As Naylor[1] notes, patients with urgent conditions may be treated quickly, even in a system with long average waiting times.

Another dynamic question that our model cannot answer is whether the waiting list will increase or decrease when the hospital's budget increases. The evidence appears to be ambiguous on this matter. After the Ontario provincial government launched a programme to increase the capacity of open-heart surgery by 800 cases per year (12% of the annual caseload), both the total waiting list and the average wait for elective surgery declined.[1] On the other hand, Harley[20] failed to detect any relationships between waiting times for trauma and orthopaedic surgery and a number of district-level measures of resource availability in the UK NHS. This finding is in agreement with our earlier observation that hospital waiting lists seem to be invariant with swings in real funding for the NHS.

In fact, these Canadian and UK findings may not be inconsistent when the circumstances surrounding the resource expansions are fully detailed. In Ontario, the provincial authorities expanded hospital capacity in a specific area, effectively pre-empting the hospitals' authority to allocate the extra funds away from coronary surgery. This was not the case in the UK data, where the budget was controlled at a micro level by the hospitals. It is not surprising that the patterns of resource allocation are different when the central authority attaches some strings to the global budget.

CONCLUSION

We have presented a model of a hospital in a global budget payment system. Patients were assumed to pay no user price for hospital services. Consequently, demand depends entirely upon patients' perceptions of hospital quality which in turn are driven by the resource intensity of services. The allocation of the hospital's budget between quantity and resource intensity is controlled by the administrators, who maximize their own utility function. We showed that excess demand for hospital services is typical of equilibrium in this model. Furthermore, an increase in the hospital's budget can lead to an increase in excess demand under plausible assumptions regarding the elasticities of demand and supply.

APPENDIX

Indifference curves between resource intensity (R) and quantity (Q)

The hospital's utility function can be written as $U = H[Q, q_{\mathrm{h}}(R)]$, where q_{h} is a function of resource intensity. We assume that $q_{\mathrm{h}}'(R) > 0$. The slope of an indifference curve is

$$\frac{\partial R}{\partial Q} = -\frac{H_Q}{H_{q\mathrm{h}} q'_{\mathrm{h}}} \qquad \text{(A1)}$$

The rate of change in the slope with respect to a change in Q is

$$\frac{\partial^2 R}{\partial Q^2} = \frac{-H_{QQ}H^2_{q\mathrm{h}} + 2H_{Qq\mathrm{h}}H_Q H_{q\mathrm{h}} - H_{q\mathrm{h}q\mathrm{h}}H^2_Q}{H_{q\mathrm{h}}{}^2 q'_{\mathrm{h}}} - \frac{q''_{\mathrm{h}} H^2_Q}{H^2_{q\mathrm{h}} q_{\mathrm{h}}{}'^{\,3}} \qquad \text{(A2)}$$

The first term in equation (A2) is the curvature of an 'ordinary' indifference curve between quality and quantity, multiplied by $1/q_{\mathrm{h}}'$. Since we assume that $q_{\mathrm{h}}' > 0$, the first term is positive provided that ordinary indifference curves are convex. The second term in equation (A2) will be zero or negative if the quality function has constant or decreasing returns (i.e. if $q_{\mathrm{h}}'' \leq 0$). Thus, the assumptions that $q_{\mathrm{h}}' > 0, q_{\mathrm{h}}'' \leq 0$ and convexity of ordinary indifference curves are sufficient to show that indifference curves between quantity and resource intensity are convex.

REFERENCES

1. Naylor, C. D. A different view of queues in Ontario. *Health Affairs* 1991; **10**: 110–28.

2. Coyte, P. C., Wright, J. G., Hawker, G. A., *et al.* Waiting times for knee replacement surgery in the United States and Ontario. *New England Journal of Medicine* 1994; **331**: 1068–71.
3. Hamilton, B. H., Hamilton, V. H. and Goldman, D. Queueing for surgery: is the US or Canada worse off?, presented at the ASSA Annual Meeting, San Francisco, CA, 6 January 1996.
4. Frankel, S. The origins of waiting lists. In: Frankel, S. and West, R. (eds), *Rationing and Rationality in the National Health Service*. London: Macmillan, 1993.
5. Government Statistical Service. Statistics of elective admissions and patients waiting: England, six months ending 31 March 1991. *Bulletin* 1991; **2**: 7.
6. Pauly, M. V. The welfare economics of medical care: comment. *American Economic Review* 1968; **58**: 531–39.
7. Buchanan, J. M. *The Inconsistencies of the NHS*. Occasional Paper No. 7. London: Institute of Economic Affairs, 1965.
8. Schieber, G. J., Poullier, J. P. and Greenwald, L. M. US health expenditure performance: an international comparison and data update. *Health Care Financing Review* 1992; **13**: 1–15.
9. Chiang, A. C. *Fundamental Methods of Mathematical Economics*, 3rd edn. New York: McGraw-Hill, 1984.
10. Feldstein, M. S. The quality of hospital services: an analysis of geographic variation and intertemporal change. In: Perlman, M. (ed.) *The Economics of Health and Medical Care*. New York: Wiley, 1974.
11. Feldman, R. and Dowd, B. Is there a competitive market for hospital services? *Journal of Health Economics* 1986; **5**: 277–92.
12. Berkowitz, E. and Flexner, W. A. The market for health care service: is there a non-traditional consumer? *Journal of Health Care Marketing* 1981; **1**: 25–34.
13. Wolinsky, F. and Kurz, R. How the public chooses and views hospitals. *Hospital and Health Services Administration* 1984; **29**: 58–67.
14. Lane, P. and Lindquist, J. Hospital choice: a summary of the key empirical and theoretical findings. *Journal of Health Care Marketing* 1988; **8**: 5–20.
15. Smith Gooding, S. K. Quality, sacrifice, and value in hospital choice. *Journal of Health Care Marketing* 1995; **15**: 24–31.
16. Phelps, C. E. *Health Economics*. New York: Harper Collins, 1992.
17. Bestard Perelló, J. J., Sevilla, F. F., Corella Muzón, M. I. and Somoza, J. E. La unidad ponderada asistencial (UPA): nueva herramienta para la presupuestación hospitalaria. *Gaceta Sanitaria* 1993; **7**: 263–74.
18. Naylor, C. D., Levinton, C. M., Wheeler, S. and Hunter, L. Queueing for coronary surgery during severe supply–demand mismatch in a Canadian referral centre: a case study of implicit rationing. *Social Science and Medicine* 1993; **37**: 61–67.
19. Lindsay, C. M. and Feigenbaum, B. Rationing by waiting lists. *American Economic Review* 1984; **74**: 404–17.
20. Harley, M. Waiting times in trauma and orthopaedic surgery. *Community Medicine* 1988; **10**: 57–65.