

CHAPTER 9

HOSPITALS CONTINUED: FROM THEORY TO MEASUREMENT

GENERALIZATIONS FROM EXPERIENCE: INDUCTION AS WELL AS DEDUCTION

Whatever the difficulties of conceptualizing hospitals as firms, it is clear that the same economic processes which go on in firms also occur in and around hospitals. Decisions are made as to what services to produce, and how, what inputs to use, and how much to pay for the inputs/charge for the services. Negotiation of a budget for reimbursement purposes involves setting implicit prices for hospital outputs, just as wage negotiation determines input prices. And concerns over hospital costs reflect concerns about the quality of decision-making in all these areas. Are the "right" sorts of hospital outputs being produced, or are people being over-hospitalized, provided with unnecessary care? Are there too many surgical procedures, unnecessary diagnostic tests, overly prolonged hospital stays? And do hospitals use the least-cost mixes of labour and other inputs in producing their services? All these are the standard questions of economic performance outlined in chapter 6, except that in health care, the issues of appropriate quantities and mixes of servicing relate back to health care needs rather than to the conventional criterion of consumer willingness-to-pay.

A substantial literature has grown up attempting to measure directly various aspects of hospital industry performance, a literature which has been at best loosely linked to theoretical models of hospital behaviour. On the whole the empirical literature has been rather more successful than the theoretical, in that it has given rise to certain generalizations about behaviour, based on statistical regularities, which do seem to have advanced our understanding somewhat. Such generalizations are of course always consistent with some implicit analytic model of hospital objectives and constraints, but unfortunately there are usually a large number of such models consistent with any particular empirical regularity.

For example, one of the most widely accepted generalizations about hospital use is Roemer's Law, that "a built bed is a filled bed." Holding all else (illness levels, incomes, insurance coverage, etc.) constant, if more beds per capita are available in a region, more will be used. (Though as noted in chapter 4, they may not be *fully* used.) And this is not because "demand" somehow defined always exceeds capacity, but because the availability of capacity modifies physicians' and hospitals' perceptions of appropriate patterns of use. Such an observed regularity, however, is consistent with a wide range of different postulates about hospital objectives.

It might appear to rule out models in which an exogenous "demand curve" specifies hospital use as depending only on out-of-pocket prices paid by users, as well as their incomes and "tastes" for hospital use, since capacity should then increase use only if it led to price cuts in a world of incomplete insurance. Roemer's Law is strongly suggestive of an agency relationship in which physicians' recommendations to patients as to hospital admissions and lengths of stay are influenced by their perceptions of bed availability. One might extend this process to include

actions by hospital managements to encourage or discourage throughput, by facilitating or obstructing early discharge, for example, or use of day care surgery.

But exogenous demand models can always be salvaged by postulating "implicit prices," unobserved variables, such that capacity lowers access costs to patients either directly or through its effects on physicians. One could argue that when bed capacity is increased, some "full price" of hospital use, of which out-of-pocket charges if any are only a part, is reduced, and hence patients choose to use more care. Unobserved variables are a powerful device to reconcile preconceived theory with awkward fact; the direct agency process by contrast has the advantage of being readily visible, even if not reported in aggregate statistics.

Empirical economic studies of hospital activities may be roughly grouped into three categories:

- (i) hospital cost function studies;
- (ii) specific hospital program studies;
- (iii) hospital utilization studies.

Cost function studies look at the hospital in aggregate, and make comparisons of performance across groups of hospitals each defined as a single unit of observation. Specific program studies focus on a particular set of illnesses, or of activities within the hospital -- "disease costing" falls in this category -- and analyse alternative ways of providing care or carrying out particular activities inside or outside hospitals. Utilization studies focus on differences in hospital use by populations compared across regions, over time, or in different systems of hospital and medical care delivery. Each type of study bears on a different aspect of the "efficiency," in the most general sense, of the hospital industry, and each has implications for the characteristics that an adequate analytic model of a hospital should have.

(i) Hospital Cost Function Studies

These studies, largely carried out by economists, have their roots in the conventional theory of the for-profit firm. Given a stable technology, expressed in a production function linking amounts of inputs of productive resources -- labour time and skills, raw materials, plant and equipment services -- with the maximum amounts of the firm's product(s) which can be produced thereby, there will exist a cost function which defines the minimum attainable cost per unit for each quantity of output produced. The short-run cost function defines minimum unit cost for given fixed capital, plant, and equipment, and hence shows unit cost rising beyond designed capacity and becoming infinite beyond absolute maximum capacity. The long-run function, however, is defined over a time horizon such that all inputs are variable and no capacity constraints apply. If plants can be replicated without increasing costs of co-ordination, then the long-run cost curve, the graph of unit cost against output, may beyond some minimum output level be horizontal -- constant returns to scale -- and unit costs will not depend on the size of the organization. Or the curve may be U-shaped, displaying a range of economies of scale -- falling unit costs as organizational size increases -- followed by diseconomies of scale as the organization becomes managerially unwieldy and unit costs rise. In the short run, of course, a U-shape is normal, as unit costs are usually elevated if a plant is being run well below or well above its designed capacity.

Whatever the shape of the cost function imposed by existing technology, the profit-maximizing, cost-minimizing firm should be operating at or near the "frontier" which it defines.

Unit costs below the cost function are, by definition, technically impossible; costs above it reflect failure to minimize costs. Thus it should be possible actually to observe the cost function for different industries by plotting unit costs against scale of operations for different firms in the same industry. One could then identify efficient firms, as well as optimal scales, and infer a number of things about market behaviour which would interest students of industrial organization.¹

The application of this statistical technique to hospitals is obvious. One could plot costs per patient-day or per admission for a group of hospitals against number of patient-days, admissions, or beds, and fit a regression line, or curve, through the scatter of points. The shape of the curve would indicate the existence or absence of scale economies or diseconomies, while outliers -- hospitals with *per diem* or per admission costs well above or below the line -- would be those with unusually efficient or inefficient managements.²

Such a relationship could be both a planning tool, in that it would show the optimal scale for building new hospitals, and a guide for reimbursement of hospitals either by arm's-length insurers or direct budget negotiators. A provincial government could squeeze the budgets of the high-cost hospitals while studying and trying to generalize the secrets of managerial success in the low-cost. And a United States insurance company or public agency might set its reimbursement rates on the basis of regional or group averages, and under-reimburse costs or charges of "above-the-curve" hospitals.

Formally, one could identify a statistical relation of the form:

$$PD_i = a + bBED_i + cBED_i^2 + e_i$$

where PD_i is the average *per diem* cost (in a particular time period) of the i th hospital in the group, BED_i is its rated bed capacity (as an index of its scale of operations), BED_i^2 allows for non-linearity -- costs may first fall and then rise as scale increases -- and e_i is the "error" specific to the i th hospital. If it is positive, the hospital has a higher *per diem* than its size warrants, and if negative, a lower, for unspecified reasons independent of the measure of size used.

Alternatively one could write:

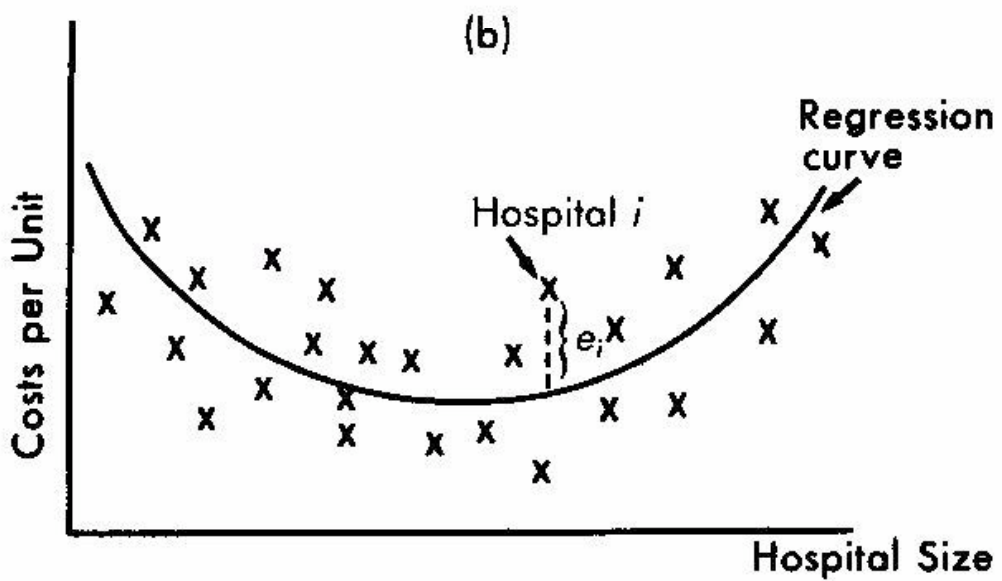
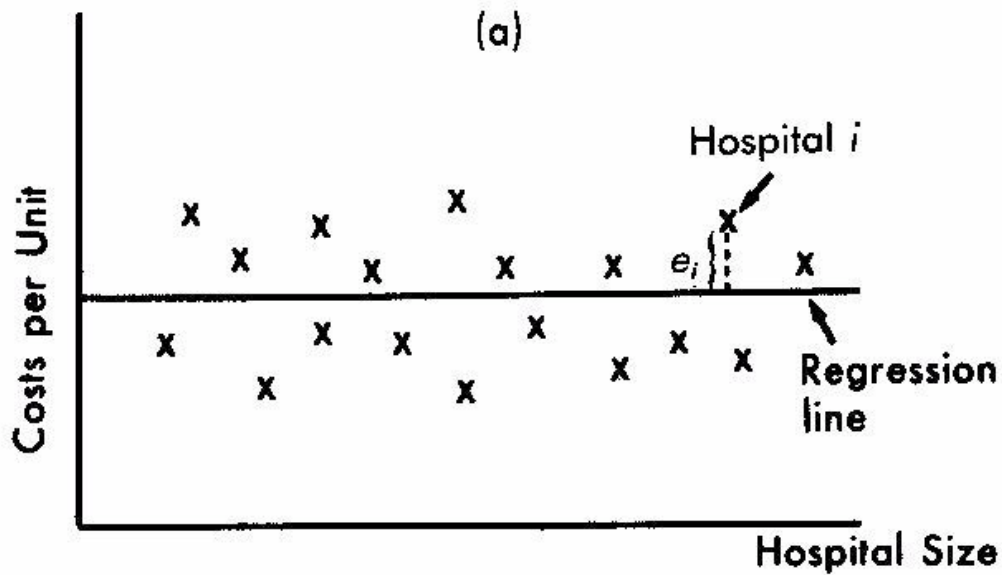
$$CC_i = a + bBED_i + cBED_i^2 + e_i$$

letting CC_i be the average cost per treated case or episode of care in hospital i . (The a , b , and c parameters are specific to each relationship, and are not constant across different equations.) CC_i would usually in practice be costs per separation (discharge or death) as an approximation to a care episode. Earlier studies tended to focus on the more familiar *per diem*, but costs per episode or at least per separation are both theoretically more satisfactory as a representation of hospital outputs -- days of care are an intermediate product used in the "production" of treatment -- and also lend themselves more naturally to statistical techniques of product standardization.

Figure 9-1 displays this cost/capacity relationship, each data point representing a single hospital, and shows outliers with unusually high and low costs. The parameters a , b , and c are constant across all hospitals in the group, and are defined by the curve of best fit (minimizing the sum of the squared e_i) to the observed scatter of points. If the relationship is one of constant returns to scale, then b and c will be zero; size does not affect cost per day (Figure 9-1a). A U-shaped relation arises (Figure 9-1b) if b is negative and c positive.

FIGURE 9-1

Hospital Costs per Unit of Output as a Function of Scale of Operation: Alternative Cost Curves



Possible Measures of:

Hospital Size: Rated Bed Capacity, Total Patient Days

Output Units: Patient Days or Separations

One can also allow for short-run effects by extending the relationship to include, for example, occupancy rates. Two dimensions no longer suffice to graph the points, but one can "plot" each hospital *per diem*, bed stock, and occupancy rate mathematically, and fit a relationship:

$$PD_i = a + bBED_i + cBED_i^2 + dOCC_i + fOCC_i^2 + e_i$$

where now OCC_i is the occupancy rate of the i th hospital. If short-run variations in occupancy influence unit costs, then one might expect d to be negative, and f positive, indicating that for any given bed stock, costs per day fall as occupancy rises (fixed costs can be spread more widely), but at some point pressure on capacity reverses this relation. If *per diem* costs do depend on both long-run scale and short-run occupancy, then the e_i in the second relation should in general be smaller than the first. The variation across hospitals would be more completely "explained" by taking account of their differences in short-run utilization as well as in scale.

Indeed, the coefficients d and f , plus knowledge of the hospital's patient-day load, enable one to calculate the marginal costs, the amount by which total costs rise, when one more patient-day of care is provided, for each size of hospital. This can then be compared to the average *per diem*. The conventional wisdom is that "an empty bed costs almost as much as a full bed," *i.e.*, given that capacity exists, the marginal costs of increased utilization are low. This would be reflected in large d (and perhaps f) estimates. Reliable estimates of marginal costs would enable provincial reimbursement agencies to determine by how much a hospital's negotiated budget should be augmented if it suffers an unexpected increase in utilization. The lower the marginal cost, the less the cost implications of short-run variations in utilization. Small values of d and f would suggest that average unit costs do not vary much with occupancy; in this case costs rise more or less proportionately with patient load.

Unfortunately, the real world is somewhat more complex than this very brief outline suggests. (For a more complete discussion, see Barer 1981, 1982). Any attempt to draw conclusions about production technology from a simple plotting of unit cost against output rests, as one might expect, on an implicit model of hospital behaviour, and not a very plausible one. It assumes that hospitals do in fact attempt to minimize unit costs, which given their not-for-profit status is at least questionable. What is in fact measured is a statistical relation between cost and output which reflects a combination of technical constraints and behavioural regularities, and shows relative costliness within a group of hospitals. It is quite conceivable that all hospitals in any group lie above the technological minimum cost curve -- if we knew where it was.

Furthermore, the measurement of costs per unit of output in terms of *per diems* or total costs per separation is badly misleading on several counts. First, if hospitals are firms, they are multiproduct firms producing some combination of inpatient care, outpatient services, education, research, and community services. Not all hospitals provide the full range, but many do, and few if any are confined to inpatient care alone. Secondly, inpatient care itself is very far from being a homogeneous process. The service intensity of a "patient-day," in terms of the diagnostic, therapeutic, and custodial services it embodies, varies widely over the range of different types and severities of problems cared for, as well as over the stages of the illness. It is obvious that some diagnoses require more intervention than others, that the severity of the "same" illness varies across patients, and that in general late days of stay for an illness episode are less costly than early. And finally, the outcome of treatment may vary with the quality of care provided, which may be connected (either positively or negatively!) with service intensity. A "treated case" of whatever type is not a measure of output uniform across hospitals if the probability of survival, or some other significant outcome dimension, varies substantially across hospitals.³

Quite apart from the problems of standardizing hospital cost data for differences in the nature of their patient load, the cost function as an instrument of management evaluation or reimbursement policy has the serious disadvantage that it assumes exogeneity of utilization. Patients appear on the hospital doorstep, are admitted, and after a time are discharged, by a process which except for absolute capacity constraints is beyond the hospital's influence. On these assumptions it is reasonable to evaluate hospitals by comparing the costs which they incur in responding to these externally generated needs or demands.

On the other hand, if hospital managements can by their administrative practices or by discussions with the medical staff influence lengths of stay, then they can influence costs per day and per separation as well. Pressures for "efficiency" focussed on *per diem* costs may encourage delays in discharge; "cheap" late days of stay serve to lower average costs. Costs per admission may be held down by admitting cases of low severity and referring high severity ones elsewhere, or by discharging and readmitting the same patient several times. Thus the simple cost comparisons may be reflecting either real differences in efficiency or just different ways of manipulating the denominators. The behavioural responses which in chapter 8 undermined the effectiveness of incentive reimbursement, may here invalidate statistical inferences.

This litany of problems has to some extent been dealt with in the course of development of "hospital cost function" studies over about twenty years. In early studies, the multiproduct problem was dealt with by inserting additional variables into the right-hand side of the PD_i relationship above, *e.g.*, to reflect the presence or absence of a medical school affiliation or to measure the numbers of outpatient visits. The Canadian hospital statistical reporting system, however, permits one to identify in each hospital the direct costs of non-inpatient activities such as outpatient care or education. By removing these, plus an appropriate share of overhead costs, one can derive an estimate of the hospital's operating (not capital) costs for inpatient care, a measure which should be more closely related to inpatient utilization than is global expenditure.⁴ Some variant of this estimate of inpatient costs has thus been used in most Canadian studies.

Variations across hospitals in the nature and severity of inpatient load have been dealt with in various ways, depending on the information available on patients. Early United States studies attempted to infer severity or complexity of illness from the intensity of servicing, and so measured surgical rates and diagnostic procedures per day or case. But this approach has the same circularity problems as the measurement of "quality" by servicing intensity described in chapter 8; a hospital which overtreats simple cases would be considered to have complex cases.

It is now generally accepted that characteristics of the patient must be used to standardize for patient load differences, since (except for misreporting) they are beyond the hospital's control. Age, sex, and diagnostic mix of patient load have been used in various transformations to try to adjust for interhospital differences. In these analyses, the separation records generated by the Canadian hospital insurance system have been extremely useful. While the specific record contents may vary across provinces, and extensive additional data are available from PAS or HMRI reports, at the very least each separation is identified by age, sex, residence, length of stay, discharge diagnoses, and operative procedure(s) if any. From these data one can develop profiles of the characteristics of each hospital's case load.

Some patient classification systems rely solely on patient age and diagnosis; others build in additional information on whether or not surgery occurred or on length of stay. In general such additional information makes systems of classification more sensitive to differences in severity within diagnostic groups, but at the cost of making a hospital's relative position, in terms of the difficulty of its case load, more sensitive to its own behaviour as well as the external needs to which it is responding. Whichever system is used, however, the adjustment process takes place in two stages. Patients are first classified into groups of differential difficulty, severity, or

complexity (labels vary), and then some aggregation rule is used to construct one or several hospital-specific measures of the difficulty of the overall case load.⁵

Certain very specific results have arisen from these studies (Barer 1981, 1982). First, the initial interest in scale economies has been seen to have been misplaced. Early United States studies gave widely divergent results (Berki 1972; Lave and Lave 1979); Canadian studies consistently showed unit costs rising with scale, more or less indefinitely, unless adjustment is made for differences in case load characteristics between small and large hospitals. After such adjustment, scale effects become rather unimportant. The characteristics of patients treated, as represented by age, sex, and diagnosis, exert a much more significant influence on relative costs, explaining half to two-thirds of differences across hospitals in costs per separation. Inpatient costs in teaching hospitals appear to be significantly higher than in non-teaching, even after deduction of costs directly allocated to education. Short-run utilization variables -- occupancy, length of stay, case flow rate (cases per bed per year) -- usually have the anticipated effects; costs per day fall as utilization rises, indicating that hospitals are mostly on the falling (negatively sloped) segment of the short-run average cost curve. Cost per episode of course rises with length of stay, but not proportionately. Marginal costs per case or per day are significant -- an empty bed is not as costly as a filled bed -- but are perhaps in the neighbourhood of half of average costs.

Most striking of all, Canadian studies consistently show that a combination of scale, short-run utilization level, and patient characteristics variables "explain" a very large proportion of inter-hospital variation in costs per day or per case -- usually in the range of 70-90 percent and in some studies close to 95 percent. This finding indicates a very high degree of managerial uniformity, presumably induced by the centralized reimbursement system and associated oversight. It does not indicate either uniform efficiency or uniform inefficiency, of course, but only that most *variations* can be "explained by," or associated with, identifiable characteristics of the hospital's situation, not differences in management.

Thus far, such aggregate statistical studies of hospitals do not appear to have been much used as input to the hospital budget negotiation process in Canada, although Quebec has applied an extension of the technique, and Alberta has some experimental work underway. Apart from the complexity of the statistical techniques, and their unfamiliarity to hospitals and to provincial negotiators, there remain some problems with the analysis.

First, variations in severity within diagnostic categories have not been allowed for, except insofar as they correlate with age and sex. One hospital's cases of diagnosis D *may* typically be "sicker" than another's. Some progress has been made with patient care classification systems in hospitals, particularly for nurse staffing, but this has not been linked to aggregate cost functions. Some United States classification systems go beyond diagnosis to include procedures performed, or direct assignments of level of severity, but as noted they are then sensitive to the hospital's choice of treatment patterns, as well as to subjective judgements of severity.

Secondly, measures of cost per "case" are based on separations -- discharges and deaths -- which are assumed to represent episodes of care. But an episode may involve re-admissions or transfers, thus generating several separations. Present studies have not adjusted for this, so its impact is unknown.

Thirdly, comparisons are made of costs per unit of *output*, not *outcome*; almost nothing has been done to test for differences in outcomes across hospitals. These in turn would represent, if adjusted for difficulty of patient load and resource availability, the differences across hospitals in quality of care. Some results do exist which indicate that mortality outcomes, at least, do not correlate with intensity of input use, whether or not one adjusts for patient characteristics. These findings cast further doubt on the identification of service intensity with "quality." But even this

limited information has not been integrated with aggregate cost function analysis (Lundman 1982).

Finally, as emphasized above, analyses at the hospital-wide level cannot deal with the issue of the appropriateness of hospital use itself. Insofar as they focus on comparisons of costs per episode, they can help to identify a combination of technical efficiency (or inefficiency) in the production of particular services, and (assuming equivalent outcomes across hospitals) more or less intensive use of resources for a given set of illness episodes. But such analyses could easily identify as highly efficient a hospital which was providing unnecessary services, admitting patients who did not need care. It is usually cheaper to look after healthy people than sick ones. The significance of such analyses depends on the extent to which one identifies general managerial efficiency or inefficiency in running inpatient services as an important policy issue in hospital care. Broader issues of over- or underuse, or of innovative techniques in substituting forms of non-institutional care, are beyond their scope.

(ii) Specific Hospital Program Studies

At a less aggregated level, a number of studies have examined the resources used and costs generated in caring for particular groups of patients in hospitals. These include comparisons across several hospitals caring for patients with the same problems, or of alternative modes of treatment in the same hospital -- day care versus inpatient surgery, radiation versus surgery for cancer -- or comparisons between hospital and non-hospital programs (home care, for example) for dealing with similar problems. The focus here shifts from the hospital to the patient or group of similar patients as the unit of analysis.

Such a study is a form of program analysis, which is discussed in more detail in chapter 11. There, however, the focus is on the methodology of such studies; in this chapter we are interested in particular results and in the implications they have for our understanding of hospital behaviour.

Studies of this type indicate that there exists substantial variation across hospitals in the process of treatment for, and costs of, similar problems. Variations shows up most readily in differences in lengths of stay; more detailed analysis shows differences in frequencies of diagnostic procedures or of surgical or other therapeutic interventions. Hospitals also differ significantly within or across regions in their use of substitutes for conventional inpatient care. Day care surgery, for example, represents a much larger proportion of total surgical load in some hospitals and provinces than in others (Evans and Robinson 1980). Care-by-parent wards for children are another, though much less widely used, form of care which has been shown to improve care quality, or at least patient/parent satisfaction, and lower costs of care (Evans and Robinson 1983). Yet implementation has been glacially slow, and highly localized.

Often the choice of program alternatives is between low and high technology interventions. A pair of widely quoted randomized trials of home care versus intensive coronary care units for survivors of a first heart attack (Mather *et al.* 1971; Hill *et al.* 1978) indicated no significant difference in outcome between the two forms of therapy -- of radically different costs. A survey of literature on the efficacy of electronic foetal monitoring (Banta and Thacker 1979) indicated that this procedure might on balance be doing more harm than good. Its high false positive rate can lead to unnecessary interventions -- Caesarian section to deliver babies "in distress" who in fact were not. This technology, widely adopted apparently without full evaluation, may be to a significant degree responsible for the current epidemic of Caesarian sections. Variations in patterns of laboratory diagnostic testing across regions and time have led to serious concerns, particularly among pathologists, as to the growing incidence of investigation without

information, of costly testing with at best zero therapeutic payoff, in hospital as well as ambulatory care.

It would be neither useful nor possible to provide here a survey of the vast array of studies on specific hospital programs. What emerges from such studies, however, is a rejection of "cost-minimizing" models of hospital behaviour. Over and over, one finds differences in techniques across hospitals which have significant cost implications without apparent associated outcome differences. Yet these differences do not lead to any response by the high-cost hospitals, at once or ever. Unlike private sector management, hospital management is not under any institutionalized pressure to seek out and adopt less costly ways of providing care.

As an example, but only an example, expert opinion in Canada holds strongly that tonsillectomy cannot safely be performed as a day surgical procedure.⁶ Although its frequency of performance has dropped dramatically in the last decade, this procedure still accounts for a very large number of paediatric admissions. Yet in British Columbia, two hospitals were each reported as having performed about a hundred such operations per year, on surgical day care, over a period of more than a decade. The important point is not whether "expert opinion" was right or wrong in this case, but that other hospitals felt no necessity to observe the discrepancy and react to it. Apparently no effort was made to find out if the "deviant" hospitals were taking undue risks, or if the conventional approach imposed undue costs.⁷

At a less local level, the serious questions raised by the randomized trial of coronary care units or the analysis of electronic foetal monitoring do not appear to have had any impact on the utilization of these technologies. The finding that halving lengths of stay for heart attack victims who have no complications in the first four days of stay had no adverse effects on patients (McNeer *et al.* 1978) has not led to drastic cuts in this major component of hospital utilization. Early discharge programs for obstetrical patients and newborns are just beginning, and struggling against apparent indifference among most physicians and hospital administrators. It is hard to see how such massive inertia in response to new technological information on ways to reduce costs while maintaining care outcomes can be reconciled with theoretical models of hospital behaviour which assume cost-minimizing behaviour a priori.

Indeed, the incentives in hospitals seem rather to be towards the adoption of more costly, more resource-intensive techniques, so long as the budgetary climate is sufficiently permissive. Far from cost-minimization, as implied by quantity-maximization under budget constraint, the process appears to be one of intensity-maximization. But the fact that much of the resulting intensity of resource use, of servicing, cannot be linked to patient outcomes, and in a number of cases can be shown *not* to be so linked, forbids us to refer to this process as "quality-maximization" -- ineffective care is not quality.⁸

Apart from their implications for our understanding of hospital behaviour, specific program studies support several important generalizations. First, there appears to be a great deal of scope for lowering hospital costs, and improving the effectiveness and efficiency of the hospital "industry," in particular program areas. No one program innovation by itself will have a major impact -- even if it were possible, say, to do all tonsillectomies on a day care basis, the influence on overall hospital use would be minimal. But across all forms of questionable utilization, the potential for reduced hospital use appears to be very large indeed. No one has yet attempted to assemble the literature on alternatives to conventional inpatient care, to see what the aggregate impact could be. A study which looked, diagnosis by diagnosis, at the savings in hospital use which have been demonstrated in some form of experimental or field trial, without deterioration of patient outcome, would almost certainly yield very large numbers indeed.⁹

Of course, hospital costs would not fall in direct proportion to utilization. A part of the savings would come from reductions in lengths of stay, and the late days of stay are those of

lowest service intensity. One should not overemphasize this point, however, as it appears that the mere fact of being hospitalized is itself a risk factor for diagnostic intervention (Hornbrook and Goldfarb 1981). Early discharge, or alternative forms of care such as surgical day care or care-by-parent for children, lead to less intensive diagnostic interventions, as well as less custodial care (Evans and Robinson 1980, 1983). And diagnostic intervention, given the inevitability of some false positives, leads on average to further care.

Reductions in hospital use lead to transfers in cost as well as reductions. People have to eat whether in hospital or out, for example I, and part of the saving of hospital dietary costs by early discharge shows up in the patient's budget. Again, however, one must not overemphasize this point. Elaborate valuations of the time and effort of patients and their families at some hypothetical market wage rate can lead to rather peculiar cost amputations by simple failure to value the benefits of not being institutionalized. In general, people strongly prefer *not* to be in hospital, and enter or stay there only because they believe they "need" care.

It is important to emphasize the extensive evidence for potential improvements in hospital efficiency, because the impression is sometimes given, particularly in economic analyses, that hospital budgeting and health budgeting generally impose a grim trade-off between cost-control and death. If all diagnostic and therapeutic interventions in hospitals (or out of them) have some expected payoff in terms of health status, then we are on the curve in Figure 1-3b, and decisions on resource allocation in hospitals are indeed decisions as to who shall live and who shall die (or who shall live in what condition). How much shall we as a society give up to extend or improve the life of someone needing care? The point of the specific program studies is that we are *not* on the curve, but well below it.¹⁰ The trade-off may be very real, and very grim, but system-wide we are not there yet. There is scope for either withdrawing resources from health care without reducing anyone's well-being, or for redeploying resources to achieve better results without having to spend more. Too much attention to the ultimate trade-off can distract our attention and our policies from these opportunities here and now.

But it is not in fact true that *no one's* well-being is reduced by changes. Patients may be better off, but providers are not. Day care surgery saves money by making ward nurses redundant. Lab tests which add nothing to diagnosis or therapy (*ex ante* or *ex post*) can add substantially to the provider's revenue, whether hospital or private tab. Since every dollar of expenditure is also a dollar of someone's income, there is a direct interest in the hospital sector in resisting cost-reducing innovation. Fee-for-service medical practice makes the conflict of interest between patient (or insurer) and provider more explicit than does salaried care, but it is present in either case. Thus it is not surprising that the second point emerging from program studies is that innovations which add to costs (incomes) proliferate rapidly on the basis of relatively weak evidence of efficacy, while innovations which lower costs (incomes) make little headway and are held to much more rigorous standards of proof of efficacy. An intervention which can clearly be shown to be harmful to patients does, of course, die out quite quickly, but those whose effects are minimal, or for which the evidence is equivocal, seem to persist indefinitely.

The failure of hospital managements to seek out new information on efficient technique, or to respond to it when it is available, is partly a problem of incentives -- not-for-profit organization and some form of cost reimbursement -- and partly a reflection of the ambiguity of the definition of management at this level. How patients are to be cared for is traditionally the prerogative of the medical staff; the administration are responsible for the efficient assembly of the resources needed to provide that care. Thus the managerial failure implicit in ineffective or unnecessarily costly modes of care is in the first instance a failure of the medical staff, individually or collectively. But the information which might guide staff decisions is generally more readily available to administration. Thus the failure of medical staff to react to new

knowledge is in part a failure of administration to assemble and present the information. In any case, the division of responsibilities in present hospital management structures appears to assign no one the responsibility for ensuring that the patterns of service which patients receive reflect the best available information on efficiency and effectiveness. And the cost implications of this failure seem much more important than the narrow questions of technical efficiency -- of cost minimization per unit of service.

Indeed a third general point arises from the specific program studies, which is even more discouraging than the problem of information dissemination and uptake. In the present climate of economic restraint, which has lasted for a decade in Canada and is probably with us for the indefinite future, new programs are very frequently advocated on the basis of cost savings. CT scanning, for example, substitutes an ambulatory procedure for several different types of inpatient procedures, and thus might save enough in inpatient care to pay its costs. Home care or day hospitals for the elderly will save money by keeping people out of institutions. Day care surgery or shortened inpatient stays can significantly reduce utilization *by particular patients*.

It does not follow, however, that overall hospital costs will in fact fall, and indeed most specific program innovations seem to be associated with increased expenditures. Two different variants of Roemer's Law are at work.

First, innovations which free up capacity of any sort induce more utilization, just as does the construction of new capacity. Hospitals frequently point out that programs to shorten stay *increase* their costs. What they mean is that *per diem* costs rise as the "cheap" late days of stay are curtailed, *and* that new admissions flow in to maintain occupancy. If admission rates did not respond, *per diem* costs would still rise, but total costs would fall. Similarly, surgical day care lowers inpatient use for the class of procedures it serves, but other forms of inpatient use react to the newly available capacity (Evans *et al.* 1983). What started off as a cost-reducing substitution of one form of care for another becomes an add-on of more care and more costs. Indeed, some United States commentators on surgical day care specifically advocate its introduction only when inpatient use is at capacity, so that one can be sure it will yield add-on business (Robinson and Clarke 1980, chapter 11). Otherwise, the new service might lower inpatient use and hence lower the hospital's revenue base!

Secondly, new procedures which substitute for more expensive (and less effective or more uncomfortable or dangerous) old procedures, as CT scanning does for pneumoencephalography, rarely stop there. Again utilization rises to meet capacity; so the utilization and overall cost of the new technique will often substantially exceed that of the techniques it replaces. Automated laboratory testing has the same result -- unit costs fall but total costs rise as volume expands. While the value of the new technique for some patients may be beyond question, at the margin it may be highly questionable. Or the new technique may simply be piggybacked on the old -- both are done "just in case."

New drug therapies show the same behaviour. Cimetidine, the H₂ receptor blocker, represented a major breakthrough in treatment of duodenal ulcer, and has been shown to improve outcomes and lower costs of care for DU cases who would otherwise have gone to surgery. Studies of its use in the field, however, show that it is being marketed and used for conditions for which its effectiveness has not been demonstrated, and is being used in conjunction with, not as a substitute for, other chemotherapy (Hall *et al.* 1981). Such widespread indiscriminate use suggests that the cost of the drug *as used* exceeds any savings of other forms of therapy it might yield; its very real effective use is surrounded by a very large penumbra of questionable and, in some cases, useless or harmful applications.

Nor is the "add-on" effect confined to technological interventions. Extended care for the elderly and chronically ill has long been promoted as a means of reducing pressure on acute care

facilities by misplaced patients, with the suggestion that overall costs could be reduced if more appropriate care were available. About twenty years of experience, however, suggests that if new facilities are built, they will be used, *and* the acute care hospitals will remain full. More recently, chronically ill or elderly patients in acute care facilities have been referred to in some quarters as "bed-blockers" -- implicitly suggesting that if they could be housed somewhere (anywhere) else, new acute admissions would be generated to use the freed space.¹¹

The extended versions of Roemer's Law, that available capacity of whatever sort induces increased utilization, are not the only sources of failure of the program substitution approach. The basic premise of substitution may in some cases simply be wrong -- homemakers and/or day hospital services may not affect people's need for and use of acute care facilities, for example, or not enough to recoup their costs. The Roemer's Law effects, however, are of particular interest. In the first place, they underscore the inadequacy of attempts to interpret hospital utilization in terms of an exogenous demand curve constraint. Hospital models which postulate such a relationship as a crucial feature will be seriously misleading, as will policy recommendations derived from such a framework. If we are to understand, model, the behaviour of hospitals, we must take explicit account of the processes whereby they influence the utilization of their own services, independently of any prices paid by patients.¹² In this context, it is obvious that "the hospital" includes its medical staff.

But secondly, the influence of capacity on use implies that policy, by whatever means introduced, must to be successful be global not partial in its impact. If substitutes for inpatient care are introduced, then a corresponding component of inpatient capacity must simultaneously be withdrawn from service. A sequence of innovations successfully carried out piecemeal will not add up to a major change in efficiency unless there exists some global constraint, public or private, over the hospital system as a whole. Hence the impossibility of effective control in fragmented, multi-source funding systems, compared with either direct public regulation of capacity in a sole source funding system such as Canada's, or the private, closed-panel, prepaid group practices in the United States which own or contract with their own hospitals.

(iii) Population-Based Hospital Utilization Studies

The third class of studies focusses on population groups rather than hospitals or specific programs, and compares aspects of hospital utilization rather than costs, efficiency, or effectiveness. Their findings, however, also serve as a basis for inferences about hospital objectives, behaviour, and performance.

Cross-population comparisons may address total hospital utilization -- patient-days and/or separations per thousand population -- or subcategories such as surgical use, or utilization patterns for particular procedures such as appendectomy or cholecystectomy. The latter are distinguished from specific hospital program studies, however, in that they would compare, say, appendectomy rates in two or more defined populations rather than patterns of service use by appendectomy patients in two or more hospitals or other provider sites.

Populations may be compared across geographical regions -- countries, large areas such as provinces or states within a country, or small areas such as counties or school districts within a province. Or they may be compared across different types of health care service organizations, as between enrollees of community health centres or prepaid group practices and users of private fee-for-service practitioner care. United States studies have also examined utilization differences between populations insured with private for-profit, private not-for-profit, and public insurance carriers.

Data assembly for such utilization studies has a number of standard problems. Definition of denominators, the population of interest, is fairly straightforward for geographic comparisons, though small-area boundaries often shift over time. But populations served by different systems of care are more ambiguous, particularly if people can as in Canada switch back and forth from, say, a community health centre to private practitioners, or can use both simultaneously without penalty. Defining utilization also poses problems, particularly for small-area studies, if use data is assembled by institution, not by patient. People cross area borders to obtain care, and one cannot assume that care provided *in* a region corresponds to care received by residents *of* the region. For small areas in Canada, boundary crossing is systematic and significant, though in regionalized delivery systems like Sweden it may be less common. For larger regions, provinces or countries, this problem becomes less severe but is replaced by definitional problems. The specific definition of a day or episode of care may vary across national statistical systems, as may the borderlines between acute hospital care, various forms of extended or chronic care, and custodial care for the elderly or disabled. Furthermore, the representativeness of aggregate or average measures of population characteristics, demographic or socioeconomic, becomes weaker as region size increases. The establishment of valid cross-population comparisons is thus not a trivial exercise.

Subject to these qualifications, a large number of comparative utilization studies have been carried out and certain strong and consistent patterns emerge.

First, there are substantial variations in hospital utilization across geographic regions at any level of aggregation. Marked variations across countries might not be surprising, as reflecting differences in population characteristics, both physiological and cultural, as well as significant differences in health care organization, delivery, and payment. But the extent of variation does not shrink as one moves to comparisons among large or small areas in individual countries. In Canada, for example, days of (public general) hospital care per capita in 1980-81 varied across the ten provinces from 25.6 percent above the national average (Saskatchewan), to 15.0 percent below (Newfoundland), and separations from +48.8 percent (Saskatchewan) to -22.1 percent (Quebec) (Canada, Statistics Canada 1982).

Surgical procedures have been most intensively studied, with comparisons between the United States and England and Wales showing utilization differences of 2:1. Canadian surgical rates exceed England by almost the same amount, while variations among the Canadian provinces for particular procedures are in the range of from 3:2 to 2:1, with some procedures as high as 3:1 (Vayda 1973; Vayda *et al.* 1975, 1976). Studies of high and low surgical regions in Manitoba show overall surgical rates about 50 percent higher in high-rate areas, while for particular procedures the ratio is from 2:1 to 3:1 (Roos and Roos 1981). In Ontario, for particular discretionary procedures, the rate per capita varied by 5:1 between the lowest and highest frequency countries (Stockwell and Vayda 1979).

Further information is, of course, necessary before one can draw inferences from such observations. It is possible that utilization differentials could correspond to different population "needs." These differentials, however, show up in age-sex adjusted data, which removes the principal correlate of hospital care "need" at the aggregate population level. And efforts to relate utilization differentials for particular procedures or diagnoses to other indicators of differential need have not in general been successful. Marked differences in length of stay for deliveries without mention of complications, for example, a well-defined procedure, hardly admit a "differential need" explanation.

Even if populations are essentially similar, or their relevant (to utilization) differences have been identified and standardized for, the observation of a differential, however large, does not in itself indicate whether one area overutilizes, or another underutilizes (or both). Additional

information is necessary, usually by particular procedure or diagnosis, as to the consequences in terms of mortality or morbidity which should follow from under- or overuse. If high use regions represent appropriate use, in terms of meeting needs/contributing to health status, then since the discrepancies are so large, low use areas should reveal significant consequences of insufficient care.¹³

In general, however, efforts to observe differences in mortality or morbidity associated with differences in hospital use have turned up very little -- except for the effects of differential use on surgical death rates. Case-fatality rates in surgery do not appear to vary with the level of utilization; apparently areas with a high rate of surgical use do not reach down to patients of lower average risk status. (Roos and Roos 1981). So deaths in surgery are proportional to the volume of surgery performed, consistent with findings that death rates fall in areas where hospitals or physicians go on strike.

It would appear that, in terms of Figure 1-3, levels of hospital utilization in Canada and probably in other countries as well are now out on the flat of the curve, beyond the point of payoff in terms of health status and perhaps even on the downward slope. And from the size of the interregional utilization differentials, we would seem to be a long way out on the flat. Substantial reductions in use would be possible in many regions without, apparently, risk to anyone's health.

If utilization differentials are not traceable to differences in population characteristics or needs and do not appear to lead to differences in population morbidity or mortality, one is led to consider explanations rooted in the delivery system itself. But simple-minded explanations in terms of physicians paid fee-for-service or hospitals maintaining their occupancy rates and revenue bases are also unsatisfactory, or at best incomplete.

There seems little question that the source of variation is in fact provider behaviour. Studies of the characteristic patterns of surgical utilization in small areas, by procedure, show stability over time but considerable sensitivity to the interests, preferences, and beliefs of the local medical community (Wennberg and Gittelsohn 1982; Roos 1983). Entry or exit of particular physician's shifts the use pattern such that Wennberg and Gittelsohn refer to the physician's "surgical signature" or personal pattern of behaviour. Efforts to relate such variations to characteristics or beliefs of the underlying population have not succeeded. The key role of specific practitioners was clearly shown in Saskatchewan when hysterectomy rates fell by nearly half in one year in response to an announcement (made to physicians, *not* the public) that performance of this procedure was to be investigated (Dyck *et al.* 1977). Even greater changes were observed in tonsillectomy rates in parts of New England in response to changes in physician information (Wennberg and Gittelsohn 1982). At the more aggregated level, Roemer's Law seems to apply to surgeons as well as to hospital beds -- where more are available more are used (Bunker 1970; Fuchs 1978).

But the relation is not a simple one. In university teaching centres, more beds and surgeons available do not appear to have the same impact, at least on common procedures, as in a non-academic environment. More generally, the effects of increased capacity, either physicians or hospital facilities, depend on the interests and preferences of the people involved and on institutional tradition and habit. There is no clear relationship, consistent across time and place, between capacity and *particular* forms of utilization.

Further evidence of the importance of the delivery system in influencing hospital use is given by the extensive information on prepaid group practices, community health centres, HMOs, HSOs, etc. in Canada and the U.S. Groups encompassing physicians and hospitals (owned or contracted) paid by capitation not fee-for-service, have been shown over and over again to generate rates of hospital utilization which are from 10 to 40 percent below those of

populations served by fee-for-service physicians. The same questions of comparability of populations and adequacy of care have been raised in response to these findings, as to the regional comparisons. But population standardization and in some cases random assignment of people to a capitation-based practice and to fee-for-service alternatives have left these findings intact. How physicians are employed and paid strongly affects how much hospital care their patients use. And despite considerable efforts to identify inadequacies of care in such practices, the reductions in use do not appear to affect mortality or morbidity. On other measures of quality, such groups frequently score above the general system.

These observations, frequently repeated over the last thirty years, could be and have been interpreted in terms of the differential economic motivations in capitation versus fee-for-service reimbursement. Not only do capitation-paid groups not profit from more servicing, they may actually gain, in a variety of ways, from keeping their patients out of hospital. But as noted above, the simple economic determinism arguments are inadequate. Inter-country comparisons have shown relatively very high servicing rates by surgeons with little or no economic stake in performance. Lichtner and Pflanz (1971) studied the German experience with appendectomy, for example, finding it three to four times the rate in any other country, with correspondingly higher surgical mortality rates, but no evidence of a higher incidence of underlying conditions. But the surgeons were salaried, not paid for the procedure. Moreover, inter-area variation within Germany was also high. McPherson *et al.* (1982) report high levels of inter-area variation in surgical rates in countries with very different funding systems. And hospital utilization rates vary across regions for capitation-based practices just as they do for private fee-for-service care. Investigators of inter-regional variations in particular forms of surgery have had mixed results in trying to find a relation with physician availability or level of activity. An important factor may be simply the uncertainty of providers about the nature of "best practice," and the evolution of personal styles or habits of behaviour, resulting from training, early experience, peer behaviour, or personal skills and references. The existence of wide variations in use, unrelated to needs or outcomes, is extensively documented, but at present there is no fully satisfactory explanation as to why.

GENERALIZATIONS ABOUT WHOSE BEHAVIOUR?

Such findings are as much, or more, a description of medical practice as of the hospital "industry." As we move from the technical efficiency with which hospitals convert resources into specific services, to the pattern of services involved in the provision of an episode of hospital care, and finally to levels of hospital use in the population as a whole, it is clear that we move from the traditional domain of the hospital administrator to that of the physician. But the necessity of this shift, if we are to think at all sensibly about the evaluation or the effective organization of the hospital sector, should be readily apparent. Hospital management includes physicians, whatever their formal organizational arrangements, in that physicians' decisions are a critical determinant of what hospitals will or will not do, as well as how they will do it. In production theory terms, the decision as to how much of what types of output to produce is principally in the hands of the physician; the administrator's influence is not negligible, but is secondary. The "flow" of production is the administrator's role. But management includes both.

On the whole, the various types of empirical studies of hospital behaviour and utilization have contributed more to our knowledge about resource-allocation processes and system performance in hospitals than have the attempts at formal modelling. They will support

generalizations which are of relevance for policy formation and evaluation, even if they are not easily embedded in some consistent conceptual model of a hospital as a behaving transactor with clear-cut objectives and constraints. Indeed, they are clearly inconsistent with the simpler forms of such models. The specific program studies make it difficult to maintain that hospitals strive to minimize costs, at least for costs *per* any measure of output meaningful to the patient or the wider society. Minimizing the costs of production of an unnecessary test is not cost-minimization. Further, the study of hospital utilization patterns and their correlates is very difficult to reconcile with the exogenous patient demand for hospital care which plays a central constraining role in many formal models of hospitals.¹⁴

But the findings are also difficult to reconcile with the Medico-Technical model of physician and hospital behaviour, or of professional behaviour generally, which assumes that there is a best way of doing everything and that the professional seeks it out on the patient's behalf. One is left with several alternative, but not mutually exclusive, hypotheses.

Physicians may recommend, and provide in hospitals, care which they know to be useless or harmful, or may deliberately choose inefficient modes of production. While this probably occurs from time to time, as in any area of human endeavour, it is difficult to believe that systematic malfeasance is sufficient to explain the discrepancies observed.

Secondly, providers may be doing what they think best, but be in error, and most importantly, have no particular incentive, and indeed positive disincentives, to seek out least cost modes of production or weed out ineffective hospital use. This seems the most plausible interpretation, leading to the question of what types of social mechanisms or organizational structures, informational or regulatory, might help to promote better (*i.e.*, more effective and efficient) performance.

Finally, one might interpret the diversity of professional behaviour as indicating that there is really no such thing as best practice or best technique, in which case it would be unsurprising that there were no consensus on it. The extreme form of this position is that medicine has nothing to do with health, so why expect hospital utilization patterns to be any more standardized than consumption of any other commodity or service? On this view, the design of curricula in professional schools would be rather difficult, to say nothing of the difficulty of justifying professional licensure and regulation. But even the advocates of such a view do not appear to take it seriously enough to follow it to its logical conclusions.

On balance, then, the most plausible interpretation of the available research seems to be that imperfections in the information available to, and incentives bearing on, the management of the hospital system -- physicians and administrators -- leads to problems of both ineffective and inefficient servicing which are quantitatively very significant. The seriousness of inefficiency in the narrower sense of resources used per procedure performed, cost per lab unit or therapeutic procedure, is less clear.

Possible institutional responses to this situation range along a continuum from efforts to stimulate efficiency through competition among hospitals in a deregulated private market-place, to a centralized public hospital service. Neither pole seems particularly attractive, or politically feasible, and certain intermediate possibilities will be considered in chapter 14, below.

NOTES

¹ There is a great deal more to the story than this, both statistically and economically, but we cannot pursue it here.

² The alert reader will notice that the concept of the production function defines the cost curve as a frontier of minimum attainable unit costs; the measurement process fits a line of best fit through a scatter of points. Beware of hobgoblins!

³ Cost comparisons among hospitals will also be distorted by differences in wage rates or other input prices faced by different hospitals. It is assumed here and elsewhere that such variations have been adjusted out, although this is by no means a trivial exercise.

⁴ Students of the wool and mutton problem will know that "joint costs are joint," and will be dubious of accountants' allocation rules as used here. Given some ideal, very large sample with plenty of observational variance, one might prefer statistical techniques for cost allocation, but in practice the accounting approach seems to yield much better results. (Judged how? A priori plausibility ...)

⁵ In Canada, such measures could be used to rank hospitals for budget negotiation purposes; in the U.S. it might appear that a global hospital measure or measures were unnecessary. If, for example, one could partition the hospital's case load into a number of mutually exclusive and collectively exhaustive categories, each believed more or less internally homogeneous with respect to its needs for treatment resources per case, then reimbursing agencies could pay each hospital the same amount, $\$X_i$, for each case in group or category i . But the relative amounts paid for different categories, the ratios X_i/X_j , would then be the aggregators yielding a single index for each hospital, which would, in fact, be its reimbursement level. And the information required to determine the X_i, X_j, X_k , does not arise from the grouping process itself. The use of Diagnosis Related Groups (DRGs) in the U.S. is intended to categorize patients insured by their federal Medicare program, but the actual rates of reimbursement will be set on the basis of average actual experience for groups of "similar" hospitals.

⁶ How often it should be performed, if at all, is a separate issue.

⁷ British Columbia Ministry of Health, Reports on Day Care Surgery (Annual), 1968 through to 1980/81. In the last five years, "expert opinion" in the U.S. has shifted to the point that day care surgery for tonsillectomy is becoming more common, a development apparently ignored by Canadian practice.

⁸ One might suggest that additional servicing can add to patient satisfaction even if it does not improve outcomes. But that depends on the service -- more back rubs may do so, but more tab tests certainly do not. It seems a safe generalization that the more technical the service, the more likely it is to lower, not raise, patient satisfaction. TLC is not high tech.

⁹ U.S. experience with HMOs represents an obvious first approximation, indicating potential savings of inpatient acute care use in the neighbourhood of 40 percent (Luft 1981).

¹⁰ Note again the ambiguity of the definition of health care "outputs." We can think of a curve of the Figure 1-3 type relating, *e.g.*, resources devoted to laboratory testing and health status, in which case we are probably on or near the curve, but it appears to be flat (or even downward sloping) where we are. Or we can think of a curve defined over resources devoted to, *e.g.*, particular forms of inpatient care, in which case unnecessary testing is represented by a point below the curve.

¹¹ Indeed, the representative of one medical association has been quoted as saying that his members' earnings are suffering because of inadequate access to acute care facilities in which to ply their trades. "Bed-blockers" are part of this problem; despite ten years of expanding capacity in extended care there is still no reason to believe that further expansion by itself would cut acute care use.

¹² Of course this process does not operate without limit; like every other "Law" it is an approximation to reality in the relevant range of experience. A hospital system such as that in the U.S. which has adapted to serving significant numbers of people who are too poor to pay their own bills or to purchase private insurance, and whose costs are therefore reimbursed by government, will find itself in severe difficulty if those subsidies are withdrawn. The hospital may still be able to keep itself full, but the clientele simply cannot pay. The only financially viable response is to move "up-market" to serve those who can afford private coverage, or who pay their own bills; but this adjustment (currently underway) cannot be instantaneous, and there may be some bankruptcies along the way.

¹³ We leave out of account here economic-theoretical explanations in terms of differences in unobserved consumer "tastes" for, *e.g.*, surgery, that might lead some populations to choose to consume more surgical operations just as they might prefer beef over mutton. Surgical procedures, like health care generally, are as noted in chapter 1 not direct arguments in the consumer's utility function -- or at least not positive-weighted ones.

¹⁴ Difficult, but not impossible. The reconciliation can be achieved with the aid of unobserved variables and circular reasoning, but the process is not particularly enlightening.